ESTADÍSTICA PARA TODOS

Estrategias de pensamiento y herramientas para la solución de problemas

Dra. Diana M. Kelmansky



Colección: LAS CIENCIAS NATURALES Y LA MATEMÁTICA

Colección: LAS CIENCIAS NATURALES Y LA MATEMÁTICA

ESTADÍSTICA PARA TODOS

Estrategias de pensamiento y herramientas para la solución de problemas

Dra. Diana M. Kelmansky

a u t o r i d a d e s

Ministerio de Educación.
Instituto Nacional de Educación Tecnológica.
Saavedra 789. C1229ACE.
Ciudad Autónoma de Buenos Aires.
República Argentina.
2009

ESTADÍSTICA PARA TODOS

Estrategias de pensamiento y herramientas para la solución de problemas

Dra. Diana M. Kelmansky



Colección: LAS CIENCIAS NATURALES Y LA MATEMÁTICA

Colección "Las Ciencias Naturales y la Matemática".
Director de la Colección: Juan Manuel Kirschenbaum
Coordinadora general de la Colección: Haydeé Noceti.

Industria Argentina

ISBN 978-950-00-0713-9

Director de la Colección:
Lic. Juan Manuel Kirschenbaum

Coordinadora general y académica de la Colección:
Prof. Ing. Haydeé Noceti

Diseño didáctico y corrección de estilo:
Lic. María Inés Narvaja
Ing. Alejandra Santos

Coordinación y producción gráfica:
Tomás Ahumada

Diseño gráfico:
Martin Alejandro Gonzalez

Ilustraciones:
Diego Gonzalo Ferreyro

Retoques fotográficos:
Roberto Sobrado

Diseño de tapa:
Tomás Ahumada

Administración:
Cristina Caratozzolo
Néstor Hergenrether

Nuestro agradecimiento al personal del Centro Nacional de Educación Tecnológica por su colaboración.

Kelmansky, Diana
Estadística para todos / Diana Kelmansky; dirigido por Juan Manuel Kirschenbaum.
- 1a ed. - Buenos Aires: Ministerio de Educación de la Nación. Instituto Nacional de Educación Tecnológica, 2009.
272 p. ; 24x19 cm. (Las ciencias naturales y la matemática / Juan Manuel Kirschenbaum.)

ISBN 978-950-00-0713-9

1. Estadística.
I. Kirschenbaum, Juan Manuel, dir.
II. Título

CDD 310.4

Fecha de catalogación: 21/08/2009

Impreso en Artes Gráficas Rioplatense S. A., Corrales 1393 (C1437GLE), Buenos Aires, Argentina.

Tirada de esta edición: 100.000 ejemplares



Dra. Diana M. Kelmansky

La Autora

Diana M. Kelmansky es Doctora en Matemática de la Universidad de Buenos Aires (UBA-1991).
Actualmente se desempeña como Profesora Adjunta en el Instituto de Cálculo de la Facultad de Ciencias Exactas y Naturales (UBA) y como Vicedirectora de la Carrera de Especialización de Estadística para Ciencias de la Salud.
Se desempeñó desde 1992 hasta 1994 como consultora del Instituto Nacional de Estadísticas y Censos (INDEC), en el marco del Programa de las Naciones Unidas para el Desarrollo (PNUD).
Desde el año 2002 hasta el 2004 fue consultora invitada por la Organización Mundial de la Salud (OMS) en el Plan de Análisis Estadístico sobre Crecimiento y Desarrollo.
Es Embajadora Educativa de la Sociedad Americana de Estadística (ASA) desde el 2005.
Dictó cursos y conferencias sobre microarrays en Argentina, México, Chile y España.
Ha publicado trabajos en revistas especializadas, nacionales e internacionales, y ha dictado numerosos cursos y conferencias en congresos abordando temáticas referidas, tanto a la estadística teórica como a sus aplicaciones en biología, economía y medicina.

ÍNDICE

1. Introducción

La **estadística** puede ser **divertida**, fácil y también **útil**.

Se la utiliza todos los días:

- Para justificar apuestas sobre el resultado de un partido de fútbol los simpatizantes comparan los rendimientos de los equipos utilizando, por ejemplo, los porcentajes de partidos ganados como local y como visitante.
- Durante la transmisión de un partido de tenis por televisión, los relatores cuentan la cantidad de tiros ganadores, puntos de quiebre aprovechados, errores no forzados, saques ganadores.
- Para diseñar pautas publicitarias, los publicistas consultan la planilla diaria de ratings (radio o televisión).
- En un mercado los consumidores observan cómo se distribuyen los precios entre los distintos puestos para realizar la mejor compra que combine calidad y precio.
- Para decidir qué alumna/o será abanderada/o de la escuela, el/la directora/a compara las notas de todos los alumnos del último año y elige el mejor promedio.

La necesitan:

- Los profesionales de la salud, para entender los resultados de las investigaciones médicas.
- Los economistas, porque cálculos eficientes les permitirán llegar al fondo de la cuestión que analizan.
- Los docentes cuando se enfrentan al problema de evaluar el rendimiento de los alumnos.
- Los sociólogos para diseñar y procesar sus encuestas.
- Los responsables de la calidad en un proceso productivo, al detectar las piezas defectuosas y controlar los factores que influyen en la producción de las mismas.
- La industria farmacéutica para desarrollar nuevos medicamentos y establecer las dosis terapéuticas.
- Los ciudadanos, para sacar sus propias conclusiones sobre los resultados de las encuestas políticas, los índices de precios y desocupación, y los resultados estadísticos que habitualmente se presentan en los medios masivos de comunicación (diarios, revistas, radio, televisión).

Muchas veces, las noticias surgen luego de varias etapas de elaboración. Sus primeros protagonistas son encuestadores, investigadores de mercado, médicos, técnicos gubernamentales y científicos de universidades o institutos. Ellos son la fuente original de la información estadística; publican sus resultados en revistas especializadas o en comunicados de prensa.

A partir de allí, entra en juego el segundo grupo: los periodistas, que pueden estar más apurados, a la caza de resultados que les permitan obtener un titular.

Finalmente, hay un tercer grupo: el de los consumidores de la información, o sea todos nosotros. Estamos frente al desafío de escuchar, leer, ver y decidir respecto a ella.

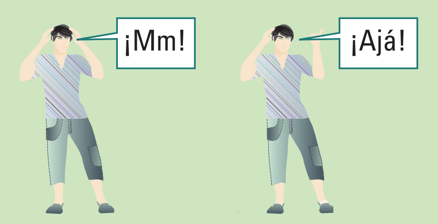
Los métodos estadísticos forman parte de cada paso de una buena investigación, desde el **diseño** del estudio, la **recolección** de los datos, la **organización** y el **resumen** de la información, el **análisis**, la elaboración de las **conclusiones**, la discusión de las limitaciones y, por último, el diseño de un **próximo estudio** a fin de dar respuesta a las nuevas preguntas que pudieran surgir.

En cualquiera de las etapas de este proceso puede haber errores. Pueden, o no, ser intencionales. **Es posible mentir con estadísticas, pero es mucho más fácil mentir sin estadísticas.**

Proponemos construir herramientas que permitan:

- Descubrir resultados engañosos.
- Obtener buenos datos.
- Distinguir entre lo que se puede y no se puede concluir a partir de una muestra.
- Entender tablas y gráficos.
- Comprender el significado de margen de error.
- Construir e interpretar intervalos de confianza.
- Tomar decisiones en base a los datos.
- Llevar a cabo estudios estadísticos sencillos

Este libro se estructura en base a ejemplos. Algunos de ellos reaparecen en capítulos sucesivos con una profundidad creciente poniendo el énfasis en el desarrollo de los conceptos. Para entenderlos y aprehenderlos hace falta **pensar**. Habrá párrafos que requerirán de varias lecturas, hasta que... "¡Eureka!", se comprende su significado.



Todos los cálculos presentados, tanto en el texto como en los ejercicios utilizan operaciones aritméticas simples, realizables con una calculadora. Debe tomarse **tiempo para pensar** las respuestas a los ejercicios sin mirar las soluciones. Estas son únicamente una guía, y para su verificación. Aunque algunas explicaciones y detalles no se dan en las soluciones estas deben formar parte de las respuestas completas a los mismos.

Utilizaremos la palabra **estadístico/a** con **cuatro significados diferentes** que, según el contexto, será fácil distinguir:

1. La **estadística como disciplina de estudio**. Siempre estará en **singular**.

2. La **estadística** o las **estadísticas** como resultados que presentan organismos de estadística oficiales como, por ejemplo, la Dirección de Estadísticas e Información de Salud -DEIS- del Ministerio de Salud y Ambiente de la Nación (http://www.deis.gov.ar/CapacitacionFetal/sistema.htm).

3. Un **estadístico** como un **procedimiento** para **obtener un número** a partir de valores de una encuesta.

4. Un **estadístico** o una **estadística** como una **persona** que tiene a la estadística como **profesión**.

De aquí que, cuando hablemos de los estadísticos o las estadísticas tendremos que ver si se trata del plural de 2, 3 ó 4.

2. Un poco de historia

Desde el antiguo Egipto hasta las actuales computadoras, pasando por los adelantados de la astronomía y del estudio de la herencia.

La **historia antigua** de la **estadística** se remonta al registro de la población que hicieron los egipcios, hebreos, chinos, griegos y romanos, desde hace unos 20 a 50 siglos. Se trata de mediciones que ya realizaba el **estado** con fines tributarios y de enrolamiento militar.

Sin embargo, las **ideas** y **herramientas estadísticas** son más recientes, surgieron lentamente de las dificultades que se plantean al trabajar con datos.

Hace dos siglos, los investigadores ya enfrentaban el problema de obtener diferentes valores **para un mismo concepto**. Debían **combinar los resultados de muchas observaciones** que, por más que las realizaran con extremo cuidado, no coincidían. Tal como sigue ocurriendo actualmente cuando, por ejemplo, se mide la altura de un niño varias veces. Se trata de la **variabilidad** debida a los **errores de medición**, cuando se obtienen resultados diferentes al medir lo mismo más de una vez.

Otro tipo de variabilidad surge, por ejemplo, en el caso de individuos de una misma población que, respecto a una misma característica, son diferentes entre sí. Por ejemplo, diferentes niños de la misma edad y género tienen distintas estaturas.

En la Argentina: El primer censo nacional se realizó en 1869. Diversos organismos tuvieron a su cargo la producción de estadísticas oficiales hasta la creación del Instituto Nacional de Estadísticas y Censos (INDEC) en 1968 (http://www.indec.gov.ar/indec/indec_historia.asp).

La Sociedad Argentina de Estadística (SAE) fue creada en 1952; es una organización técnico científica, sin fines de lucro, dedicada a promover el desarrollo de la Estadística en nuestro país.

Hacia comienzos del siglo XIX, los **astrónomos** utilizaban en forma generalizada **métodos estadísticos** y escribían **textos** razonablemente **sencillos** para explicarlos. Para describir la variabilidad de sus observaciones, resultantes de los inevitables errores de medición, utilizaban como modelo matemático la distribución Normal o Gaussiana, porque les permitía explicarla con solo dos valores: la media y el desvío (Se verá con más detalle en los Capítulos 17, 18 y 20).

La distribución Normal también se utilizó para caracterizar la variabilidad entre individuos de una población, respecto de alguna característica, como por ejemplo, el perímetro cefálico. Ya no se trata de una **variabilidad** debida a los errores de medición, sino a las **diferencias entre un individuo y otro.**

Originalmente, la **estadística** estuvo limitada al cálculo de **medidas resumen.** Por esa razón, existe una directa asociación entre "hacer una estadística" y "calcular un promedio o un porcentaje". Esta última, es la estadística que encontramos habitualmente en los medios de comunicación: **promedios, porcentajes, gráficos de barras.**

Pero la estadística es más que el cálculo de promedios y porcentajes. Específicamente, se trata de hallar el rango de valores dentro del cual pueden encontrarse los datos o la mayoría de ellos, es decir, caracterizar su **variabilidad** y, más generalmente, su **distribución** completa. Conocer la distribución de los datos es importante. Un **promedio** puede tener **significados muy diferentes** según sea la forma en que se distribuyen los valores.

Algunos realizaron estimaciones con singular éxito: entre julio y septiembre de 1882, sí, ¡1882!, el astrónomo y matemático, canadiense/norteamericano, Simon Newcomb logró una estimación bastante precisa (299.810 km/s) de la velocidad de la luz realizando una combinación ponderada de una serie de observaciones. Actualmente, se considera que la velocidad de la luz en el vacío es 299.792,458 km/s, o sea aproximadamente 300 mil km/s.

Otros, en cambio, sin éxito: en 1869, el economista inglés, William Jevons, fue acusado de combinar mal los precios de diferentes productos en un índice para estudiar las variaciones del precio del oro. Los índices de precios fueron y seguirán siendo siempre un gran dolor de cabeza.

Cuando se quiere obtener conclusiones respecto de **toda la población** pero no es posible, o no es deseable, registrar datos de esa población completa, se los obtiene de algún subgrupo o **muestra de la población.** Este proceso se denomina **inferencia estadística.**

La **inferencia estadística** como disciplina nació en la primera mitad del Siglo XX con el surgimiento de los **diseños estadísticos** para obtener datos y el desarrollo de métodos para analizarlos. Sin embargo, fueron los últimos 30 años, en especial con el advenimiento de las computadoras, los que vieron la explosión de su desarrollo y aplicación.

La invasión del **mundo digital** a nuestras vidas (computadoras, acceso a Internet, teléfonos celulares, cámaras digitales) acelera los procesos de obtención y difusión de la información. Todos los campos de estudio ponen **mayor énfasis en los datos.**

La estadística se ha transformado en un método central del conocimiento. Toda **persona educada** debería estar familiarizada con los **conceptos estadísticos.**

3. Los datos son noticia

Cada mañana nos enfrentamos con una gran cantidad de información estadística que abarca prácticamente todo: desde deportes, política y economía, hasta los avisos publicitarios.

Presentaremos varios ejemplos reales que ilustran esta situación, e iniciaremos una línea de análisis, que ampliaremos en los capítulos siguientes, para determinar si las conclusiones que presentan son las adecuadas.

□ 3.1 Encuestas de opinión

Una encuesta de opinión es un mecanismo para acercarse a la visión que tiene el conjunto de la sociedad, o algún subgrupo, sobre un determinado tema. Se utilizan diferentes métodos: preguntas en la calle, por teléfono, mediante citas previas, etc. Generalmente, requieren respuesta voluntaria.

Ejemplo: Un diario presenta los resultados de una encuesta de opinión.

Se había consultado a la gente si los **mensajes del correo electrónico** deben ser contestados de inmediato.

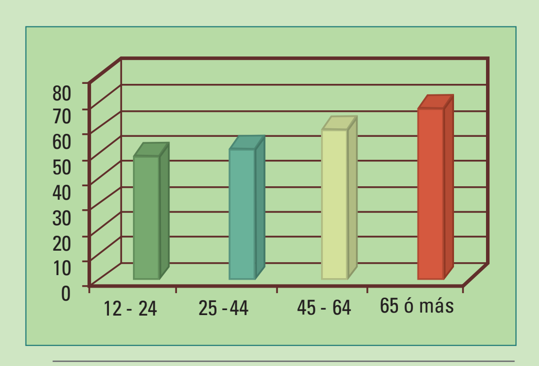


Fig 3.1. *Porcentaje de respuestas a favor de contestar los mensajes del correo electrónico en forma inmediata por grupo de edad. Fuente: USA Today 19 Ago 2008.*

El diagrama de barras muestra los porcentajes de respuestas afirmativas por grupo de edad (12-24, 25-44, 45-54, 65 ó más años). Los porcentajes obtenidos fueron 53%, 54%, 61%, 71% respectivamente. El artículo en el que se muestran estos datos concluye: **"Los mayores responden más rápido el correo electrónico".**

¿Es una conclusión adecuada? ¿Se trata de una respuesta a otra pregunta?

¿Se vincula la conclusión con la pregunta planteada?

Habría que saber un poco más sobre cómo se hizo la encuesta:

1. ¿Cuántas personas respondieron en cada grupo? Si la cantidad de personas de los grupos de edades es muy diferente, los porcentajes (los estadísticos) calculados en cada uno de ellos tiene diferente confiabilidad, y podría no tener sentido compararlos.

2. ¿Cómo fueron elegidas esas personas? Podrían ser las primeras 100 en la puerta de una escuela.

3. ¿Qué características tienen los consultados? Podrían ser todas mujeres.

4. ¿Representan esas personas a la mayoría de la población con su misma edad como para concluir que los mayores responden más rápido el correo electrónico? Si las personas encuestadas fueron las 100 primeras personas que salieron de la escuela es muy posible que sus características no representen a la mayoría de la población.

5. ¿Los usuarios de Internet mayores de 65 años son como el resto de sus contemporáneos o son distintos? Es posible que los usuarios de Internet que tienen más de 65 años tengan diferentes inquietudes que los hombres y mujeres de su misma edad y no sean usuarios de Internet.

6. ¿Opinar que un correo electrónico debe responderse de inmediato, es lo mismo que efectivamente hacerlo? La pregunta de la encuesta se refiere a si los mensajes del correo electrónico **deben ser** contestados de inmediato. Eso no implica que los que contesten por sí, necesariamente lo hagan. Por lo tanto, la conclusión "Los mayores responden más rápido el correo electrónico", **no puede obtenerse a partir de la encuesta realizada.**

◻ 3.2 Publicidad

Veremos dos ejemplos que aparecen habitualmente en los medios de comunicación. Se apoyan en resultados estadísticos, sin embargo, sus conclusiones no se ajustan a los mismos.

3.2.1 Crema reductora

La **publicidad de una crema reductora** afirma: ¡3 cm menos!

En letra chica dice:

*"Primeros resultados medibles en muslos, caderas, panza y cintura luego de 4 semanas de uso** 80% de mujeres convencidas*
Reducción de hasta 3cm en el contorno de muslos, caderas, panza y cintura entre las 4 y 8 semanas de uso
***Testeado en 168 mujeres durante 3 semanas"*

¿Se encuentra una justificación suficiente a la afirmación ¡3 cm menos!, en letra chica? Veamos.

- ¿Qué significa una reducción de "hasta 3 cm"? Significa que como máximo se obtendrá una reducción de 3 cm, pero puede ser menor. Aunque no ocurriera una reducción, o

incluso si se diera un aumento de la cintura no nos estarían mintiendo.
- Una información más útil podría ser el rango de valores obtenidos. No es lo mismo que las reducciones se encuentren entre 0 y 3 cm, a que estén entre 2,5 y 3 cm. En el primer caso podría haber muchas personas a las que la crema no les hizo absolutamente nada, y en el segundo, la crema parece haber sido efectiva para todos.
- Además, la publicidad sugiere un uso de entre 4 y 8 semanas, para ver si logramos un resultado, cuando la crema fue testeada durante 3 semanas. ¿Cómo pudieron llegar a esa conclusión?

Muchas veces en los medios de difusión, como ocurre en este ejemplo, **se refuerza una afirmación con argumentos estadísticos falsos.**

3.2.2 Pasta dental

La publicidad de una pasta dental afirma que 4 de cada 5 odontólogos recomiendan una marca determinada. ¿Cuántos dentistas fueron encuestados? No se sabe. Porque la publicidad no lo dice. ¿Por qué importa saber la cantidad de respondentes? La fiabilidad del resultado depende de la cantidad de información que se analice, siempre que ésta sea de buena calidad (veremos en los próximos capítulos cómo se produce información de buena calidad).

Cuando los anunciantes dicen “4 de 5 odontólogos” es posible que en realidad hayan sido 5 los odontólogos encuestados, o ninguno si es que inventaron el resultado. También pueden haber sido 5.000 y 4.000 recomendaron dicha marca, que no es lo mismo. No se sabe cuántos dentistas realmente recomiendan esa pasta.

□ 3.3 Razón, tasa y porcentaje

Los estadísticos que se utilicen para describir cantidades pueden hacer una diferencia, respecto a las conclusiones que se obtienen. Primero, veremos algunas definiciones para, finalmente, desarrollar un ejemplo sobre la medición de los accidentes de tránsito.

3.3.1 Definiciones

Una razón es el **cociente entre dos cantidades.** Por ejemplo, “La razón de niñas a niños es de 3 a 2” significa que hay 3 niñas por cada 2 niños. No debe entenderse que sólo hay 3 niñas y 2 niños en el grupo. Las razones se expresan utilizando los términos más bajos para simplificar lo más posible. Así, esta razón expresa la situación de un curso de 25 alumnos con 15 niñas y 10 niños, o de un colegio con 300 chicas y 200 chicos.

Una tasa (o velocidad) es un **cociente que refleja una cierta cantidad por unidad.** Por ejemplo, un automóvil se desplaza a 45 km por hora (la unidad es una hora), o la tasa de robos en un barrio, 3 robos por cada 1.000 hogares (la unidad es 1.000 hogares).

Un porcentaje es un **número entre 0 y 100** que mide la **proporción** de un total. Por ejemplo, cuando decimos que una camisa tiene un 10% de descuento, si el precio original (el total) es $ 90, el descuento es de $ 9. Si decimos que el 35% de la población está a favor de un período de cuatro días de trabajo a la semana, y la población tiene 50.000 habitantes, entonces son 17.500 (50.000 x 0,35 = 17.500) los que están a favor. La proporción de los que están a favor es 0,35.

- Un porcentaje del 35% es lo mismo que una proporción de 0,35
- Para convertir **un porcentaje** en **una proporción,** se **divide** al porcentaje por 100.
- Para convertir una proporción en un porcentaje, se multiplica la proporción por 100.

3.3.2 Variaciones relativas

Cuando un porcentaje se utiliza para determinar un aumento o reducción relativa (relativa al valor inicial), se denomina **variación porcentual.**

Supongamos que la cantidad de accidentes por año en una ciudad pasó de 50 a 60, mientras que la cantidad de accidentes en otra ciudad pasó de 500 a 510. Ambas ciudades tuvieron un **aumento** de 10 accidentes por año, pero para la primera ciudad, esta diferencia como porcentaje del número inicial de accidentes, es mucho mayor.

Variación porcentual: se toma el valor "después de" y se le resta el "antes de", luego se divide ese resultado por el "antes de". Así, se obtiene una proporción. Para transformarla en un porcentaje se multiplica el resultado por 100.

Para la primera ciudad, esto significa que la cantidad de accidentes aumentó en un

$$\frac{60 - 50}{50} = \frac{10}{50}$$

$$= 0{,}20 \text{ ó } 20\%$$

Para la segunda ciudad, este cambio refleja sólo un aumento del 2%, pues

$$\frac{510 - 500}{500} = \frac{10}{500}$$

$$= 0{,}02 \text{ ó } 2\%$$

Si una ciudad pasó de 50 a 40, mientras que en otra la cantidad de accidentes pasó de 500 a 490, ambas ciudades tuvieron una **reducción** de 10 accidentes. Calculemos las variaciones en este caso:

$$\frac{40-50}{50} = \frac{-10}{50} = -0{,}20 \text{ ó } -20\% \quad \text{y} \quad \frac{490-500}{500} = \frac{-10}{500} = -0{,}02 \text{ ó } -2\%$$

Las reducciones se reflejan en variaciones porcentuales negativas.

Las variaciones relativas se pueden expresar como variaciones porcentuales o proporciones.

3.3.3 ¿Cantidades o tasas?

El resultado puede ser diferente según que estadístico se elija. Veamos un ejemplo.

3.3.3.1 Accidentes de tránsito

¿Cómo medimos los accidentes de tránsito? Veamos dos maneras de analizar las estadísticas sobre los accidentes de tránsito, mostrando dos aspectos diferentes de la misma historia.

Muchas veces el análisis puede utilizarse con fines políticos. Un candidato puede argumentar que los accidentes fatales se han reducido durante su mandato y su contrincante que han aumentado. A partir de una misma realidad, ¿cómo pueden los dos candidatos decir que la cantidad de accidentes fatales evoluciona en dos direcciones diferentes?

Consideremos los datos de la tabla 3.1, que muestran la cantidad total de víctimas mortales por accidentes de tránsito en el lugar del hecho, en la Argentina desde el año 2.000 hasta el año 2007, de acuerdo con el Registro Nacional de Antecedentes de Tránsito (R.e.N.A.T.).

La cantidad de víctimas mortales se redujo desde el año 2.000 hasta el 2004. A partir de ese año, los accidentes comenzaron a aumentar. Podría decirse que en el 2007 estuvimos peor que en el 2001, con 135 muertes más (4.175 contra 4.040). Pero, ¿**la cantidad de víctimas mortales** es la medida **adecuada** para describir el problema?

Estas cifras no dicen toda la historia. Una parte importante de la información ha quedado fuera. Aumentó la cantidad de accidentes fatales, pero también aumentó la cantidad de vehículos circulantes. Ante iguales condiciones de conducción, es razonable esperar que si aumentan los vehículos circulantes aumentarán los accidentes de tránsito y, por lo tanto, las víctimas fatales. Para poner el problema en perspectiva es necesario incluir en el análisis, tanto la cantidad de vehículos circulantes como la cantidad de muertes. ¿Cómo se hace?

El Registro Nacional de Antecedentes de Tránsito (R.e.N.A.T.) publica además de la cantidad de muertes, la tasa de muertes por cada 100.000 vehículos en circulación.

VÍCTIMAS MORTALES POR ACCIDENTES DE TRÁNSITO EN EL LUGAR DEL HECHO EN LA REPÚBLICA ARGENTINA (2000 - 2007) TABLA 3.1

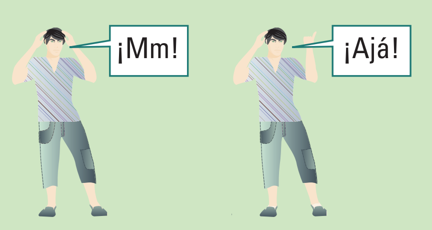
Año	Cantidad total muertes	Cantidad de vehículos circulantes	Tasa c/100.000
2000	4.316	6.799.114	63,48
2001	**4.040**	6.937.355	**58,24**
2002	3.830	7.005.406	54,67
2003	3.690	7.102.855	51,95
2004	3.047	7.355.731	41,42
2005	3.378	7.717.513	43,77
2006	3.842	7.923.726	48,49
2007	**4.175**	**7.995.043**	**52,21**

Fuente: *http://www.renat.gov.ar/Estadistica.htm*

Comparando las tasas de muertes, nuevamente entre los años 2007 y 2001 vemos que se redujo (52,21 contra 58,24).

Una tasa es un cociente. Refleja una cantidad dividida por una cierta unidad. Por ejemplo una velocidad, espacio / tiempo (km/hora), es una tasa.

¿Cómo dijo?



¿Qué significa una tasa de 52,21 **muertes** por accidentes **cada 100.000 vehículos?** 52,21 es la cantidad de muertes y la unidad en este caso es 100.000 vehículos.

¿Cómo se obtiene?

$$\frac{\text{cantidad de muertes}}{\text{cantidad de vehículos}} \times 100.000 = \frac{4.175}{7.995.043} \times 100.000$$

$$= 52,21$$

Tenemos que multiplicar por 100.000 porque $\frac{\text{cantidad de muertes}}{\text{cantidad de vehículos}}$ es la cantidad de accidentes que le corresponderían a un vehículo.

¿Cantidades o tasas?

Dependiendo del estadístico utilizado para resumir la información, para este caso cantidades o tasas, se pueden obtener conclusiones opuestas, como las que obtuvimos al comparar los años 2001 y 2007 en relación a los accidentes fatales.

¿Cuál es el estadístico correcto? Depende.

"Aquí estamos utilizando el término estadístico con dos significados diferentes:

1. procedimiento para obtener un número
2. disciplina"

Muchas veces la respuesta en un ámbito **estadístico** es: depende.

Depende de la pregunta que queramos responder. Si nos interesa evaluar el éxito de la política de educación vial deberíamos utilizar tasas de muertes; pero si organizamos el servicio de ambulancias importa la cantidad de accidentes y no los motivos (aumento de vehículos, aumento de la población, aumento de la cantidad de conductores imprudentes).

En el 2007 tuvimos **4.175** víctimas mortales por accidentes de tránsito, esta cantidad de muertes equivale a la caída de un avión jet sin sobrevivientes cada quince días (un avión jet lleva aproximadamente 150 pasajeros). ¡Eso es mucho!

3.4 Actividades y ejercicios

1. El 7 de septiembre de 2008 podía leerse en un diario: "El producto bruto de Brasil es 1.313 billones de dólares. Es cuatro veces superior al PBI que hoy tiene la Argentina". ¿Le parece adecuada esa conclusión teniendo en cuenta que la población de Brasil era para esa fecha 189,3 millones de habitantes, mientras que la nuestra era de 39,7 millones?

2. Un investigador señala que la cantidad de accidentes en su ciudad es mayor entre las 18 h y 20 h (tarde tarde) que entre las 14 h y las 16 h (tarde temprana). Concluye que la fatiga juega un papel muy importante en los accidentes de tránsito, porque los conductores están más cansados durante la tarde tarde que durante la tarde temprano. ¿Considera que esta conclusión está bien justificada?

3. Halle 3 ó más **noticias** o artículos de opinión que presenten, tasas, proporciones o porcentajes (o algún otro cálculo de tipo estadístico) para justificar un punto de vista.

4. Halle 3 ó más avisos publicitarios que muestren resultados de estudios estadísticos para resaltar la efectividad o preferencia de un producto.

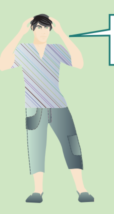
5. Realice las preguntas que considere necesarias para evaluar las siguientes afirmaciones de un aviso publicitario anunciando un producto contra la celulitis:

- En 15 días piel de naranja menos visible*
- Piel más lisa 86%**
- -1,9 cm en 4 semanas*

**Test clínico en 50 mujeres. **Autoevaluación sobre 44 individuos.*

6. Explique las siguientes frases:

- Le puedo pagar a lo sumo $ 500 por ese trabajo.
- Le voy a pagar como mínimo $ 500 por ese trabajo.
- Quiero que vuelvas como máximo a las 11 de la noche.
- Se presentaron por lo menos 10 personas para el puesto de encargado de control de calidad.
- No más de 10 personas se presentaron para el puesto de chofer.



7. "Un chico de 8 a 12 años puede perder hasta un litro de transpiración durante dos horas de actividad un día caluroso", afirma una publicidad. Nos preguntamos:

- ¿cómo se podrá llegar a esa conclusión?
- ¿cómo será para los de 13 a 16 años?
- Aunque sea complicado, proponga algún procedimiento para estimar cuanto líquido puede perder un chico por transpiración durante dos horas.
- "Hasta dos litros" ¿significa que puede:
    - no perder nada?
    - perder 3 litros?
    - perder 2 litros?
    - perder 1 litro?
    - perder 1,5 litros?
    - perder 2,5 litros?

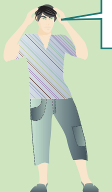
4. Herramienta para la ciencia

La estadística interviene activamente en todas las etapas que componen el método científico.

Aunque para el método científico no exista una secuencia única, señalamos los siguientes pasos generales:

- planteo de preguntas,
- planificación y realización de los estudios,
- recolección de datos,
- análisis de la información,
- obtención de las conclusiones.

Aunque en sí misma es una ciencia que se dedica al desarrollo de nuevos métodos y modelos, la estadística es una disciplina que ofrece un conjunto de ideas y herramientas para el tratamiento de datos. En este sentido, podemos decir que se trata de una disciplina metodológica, su necesidad se deriva de la omnipresencia de la variabilidad.



Los estadísticos trabajan junto a expertos en diferentes disciplinas.

La estadística está involucrada en el proceso que va desde la recolección de la evidencia, el planteo de las preguntas, hasta hallar las respuestas.

Lo importante es que haya preguntas. Para hallar respuestas se pueden seguir diferentes caminos.

Toda investigación comienza con preguntas como:

- ¿Hace mal comer papas fritas?
- ¿Cuánto cuesta enviar un/a niño/a al colegio?

- ¿Quién ganará las próximas elecciones de un club de fútbol?
- ¿Está el peso del cerebro, relacionado con la inteligencia?
- ¿Es posible tomar demasiada agua?

Aunque ninguna de las preguntas anteriores se refiere directamente a números, para responderlas se requiere del uso de datos y de un procedimiento estadístico.

Veamos un ejemplo:

Supongamos que una investigadora quiere determinar quién ganará las próximas elecciones para presidente del Club Grande de Fútbol (58.210 socios), y supongamos también que ya tiene la pregunta, para responderla deberá seguir los siguientes pasos:

- **Determinar el grupo de personas a participar del estudio**

  Puede utilizar la lista de socios en condiciones de votar.

- **Recolectar los datos**

  Este paso es más difícil. No se puede preguntar a todos los socios si van a ir a votar, y en el caso afirmativo, ¿a quién? Supongamos que alguien dice que irá a votar y declara a quién votará: ¿realmente irá esa persona a votar el día de las elecciones?; ¿dirá esa persona a quién piensa votar realmente? ¿Y si el día de la votación cambia de opinión?

- **Organizar, resumir y analizar los datos**

  Luego de recolectar los datos, la investigadora necesita organizar, resumir y analizarlos. Habitualmente, esta parte de su trabajo se reconoce como una tarea para estadísticos.

- **Obtener los resúmenes, tablas y gráficos, realizar el análisis, conclusiones tratando de responder la pregunta del investigador.**

  Presentaremos los tipos de resúmenes, tablas y gráficos que podrá utilizar a partir del capítulo 7. Aunque no podamos todavía describir detalladamente el análisis, sabemos que la investigadora no podrá tener un 100% de confianza en que sus resultados sean correctos, porque no le ha podido preguntar a todas las personas y, además algunas pueden cambiar de opinión. Pero sí es posible tener una confianza cercana al 100%, digamos 95% de que la estimación es correcta. De hecho, si ha tomado una muestra representativa (sección 5.4.1) de –por ejemplo- unas 600 personas, de manera que todos los socios tienen igual chance de ser elegidos (muestra insesgada), puede tenerse un resultado preciso con un margen de error de más o menos 4%. Siempre existe la posibilidad de que la conclusión de un estudio sea errónea. Veremos más adelante (sección 10.1) que **el margen de error sólo depende del tamaño de la muestra** y no del tamaño de la población. También veremos el significado del porcentaje de confianza y cuán preciso se espera que sea un resultado.

- **Nuevas preguntas**

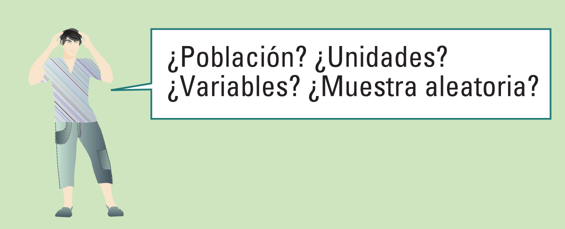
  Cuando se concluye una investigación y se han contestado las preguntas, los resultados suelen llevar a nuevas preguntas. Podría averiguarse porqué los socios jóvenes prefieren al candidato Rolando Forzudo y los socios mayores a su oponente. A Forzudo podría interesarle estudiar qué factores hacen que los jóvenes realmente vayan a votar.

Hemos dicho que la necesidad de la estadística se deriva de la omnipresencia de la variabilidad. Pero, ¿dónde se encuentra la variabilidad en el ejemplo de la encuesta sobre la preferencia del candidato? Hay varias fuentes (o motivos) que producen variabilidad. La primera, y la más importante, es que no todos los socios piensan igual, si lo hicieran alcanzaría con saber que piensa uno de ellos. La segunda resulta porque las personas cambian de opinión, si esta fuente de variabilidad es muy grande, la validez del resultado depende de cuán cerca de las elecciones se realice la encuesta. La tercera se debe a que los encuestados pueden mentir (sección 6.3.2.).

5. Vocabulario - jerga

Los conceptos requieren de palabras específicas para ser identificados.

La estadística tiene su propio vocabulario. Veremos algunos términos básicos, que volveremos a encontrar más adelante, además, seguiremos incorporando términos a lo largo de todo el libro.



Con la intención de fijar ideas, retomemos la investigación para saber quién ganará las próximas elecciones como presidente del Club Grande de Fútbol.

El primer paso es determinar el **grupo de personas a ser estudiadas,** o sea determinar la **población en estudio.** En este caso es la totalidad de los socios del Club Grande de Fútbol en condiciones de votar.

El segundo paso es **recolectar los datos.** Aquí aparecen varias cuestiones que nos permiten ilustrar más términos específicos. ¿Cuáles **individuos** serán encuestados?, esto es, ¿cuál será la **muestra**? ¿Se los elegirá en forma **aleatoria** de manera que todos los socios tengan la misma oportunidad de ser seleccionados? ¿Qué **variables** (edad, género) serán importantes en relación al tema central de la encuesta (candidato preferido)?

◻ 5.1 Unidades muestrales

A los objetos de interés de un estudio se los denomina **unidades muestrales** o simplemente unidades. Muchas veces, las unidades muestrales son **individuos**: tornillos, personas, tubos de pasta dentífrica, lamparitas. Otras veces, las unidades están compuestas por **muchos individuos:** ciudades, escuelas, lotes (de tornillos) etc.

5.2 Variables

Las **variables** son **características** que pueden **cambiar** de una **unidad** muestral a otra, como la **edad** de las personas, la población de cada ciudad, el **porcentaje** de alumnos reprobados de una escuela, la **preferencia** de una comida balanceada para un animal, la **intensidad** de emisión de rayos X de cada televisor, la **capacidad** de almacenamiento de un disco rígido, la **longitud** de un tornillo, la **duración** o el consumo de una lamparita.



No confundir una **unidad** muestral como **objeto** completo y diferenciado que se encuentra dentro de un conjunto (una docena tiene doce unidades) con las unidades que se utilizan para valorar una magnitud (el metro es una unidad de longitud).

5.3 Población

Para cualquier pregunta que interese responder, primero es necesario dirigir la atención a un **grupo particular de unidades** muestrales: personas, ciudades, animales, televisores, discos rígidos, tornillos o lamparitas.

- ¿Qué piensan los porteños sobre el Sistema de Evaluación Permanente de Conductores?
- ¿Qué porcentaje de familias de la ciudad de Santa Fe tienen mascotas?
- ¿Cuál es la expectativa de vida de los diabéticos?
- ¿Qué porcentaje de todos los tubos de pasta dentífrica son llenados de acuerdo a sus especificaciones?
- ¿Cuál es la duración promedio de las lámparas de bajo consumo de una determinada marca?
- ¿Los jóvenes deportistas consumen menos alcohol que los sedentarios?

En cada uno de los ejemplos, se plantea una pregunta y se puede identificar uno o más grupos específicos de unidades que interesa estudiar: los porteños (habitantes de la ciudad de Buenos Aires), las familias de la ciudad de Santa Fe, los diabéticos, los tubos de pasta dentífrica, las lámparas de bajo consumo, los deportistas y los sedentarios.

Se llama **población a todo el grupo de unidades muestrales** (generalmente son individuos) que interesa estudiar con el fin de responder una pregunta de investigación. Las poblaciones, sin embargo, pueden ser difíciles de definir. En un buen estudio, los investigadores deben **definir la población con toda claridad.**

La pregunta respecto a si los jóvenes que practican deportes consumen menos alcohol, sirve de ejemplo para ver lo difícil que puede ser definir con precisión la población. ¿Cómo definir un joven? ¿Los menores de 18 años de edad? ¿Los menores de 30 años? ¿Cómo definiría un sedentario? ¿Interesa estudiar los jóvenes de la República Argentina

o los de todo el mundo? Los resultados pueden ser diferentes para los menores de 18 que para los mayores, para los latinoamericanos comparados con los europeos, y así otras clasificaciones.

Muchas veces, los investigadores quieren estudiar y sacar conclusiones sobre una **población amplia** pero, con el fin de ahorrar tiempo, dinero, o simplemente porque no se les ocurre nada mejor, sólo estudian una **población muy restringida.** Esto puede conducir a serios problemas al momento de sacar conclusiones.

Supongamos que un profesor universitario quiere estudiar si los jóvenes que practican deportes consumen menos alcohol. Basa su estudio en un grupo de sus alumnos, que participan porque al hacerlo se les da cinco puntos adicionales en su puntaje final. Este grupo de alumnos constituye una muestra; pero los resultados no pueden generalizarse a toda la población de jóvenes, ni siquiera a todos los estudiantes.

☐ 5.4 Muestra

¿Qué hacemos para probar la sopa? Revolvemos la olla con una cuchara, sacamos una porción -una muestra- la saboreamos y sacamos una conclusión sobre toda la sopa de la olla sin haber en realidad probado toda. Si la muestra ha sido tomada adecuadamente -sin elegir tramposamente la parte buena- tendremos una buena idea del sabor de la totalidad de la sopa. Esto se hace en estadística, más específicamente en **inferencia estadística.**

Los investigadores quieren averiguar algo sobre una población, pero no tienen tiempo o dinero para estudiar a todos los individuos que la conforman. Por lo tanto, ¿qué hacen? Seleccionan una **cantidad pequeña de unidades muestrales de la población** (esto se llama una **muestra**), estudian esas unidades, generalmente individuos, y utilizan esa información para sacar conclusiones sobre toda de la población.

5.4.1 Muestra representativa

Nos interesa obtener **"buenas muestras".**



Una buena **muestra** debe ser **representativa** de la población. Esto significa, que todas las características importantes de la población tienen que estar en la muestra en **la misma proporción que en la población.**

Una muestra tiene, en pequeño y lo más parecidas posibles, las características de la población.

Podremos sacar conclusiones respecto de la población total a partir de una muestra -esto es, realizar una inferencia-, para todas aquellas características en las cuales la muestra representa a la población.

Ejemplo:

Consideremos un ejemplo simple, una población constituida por personas que difieren entre sí en una única característica con dos categorías:

Característica: el peso
Categorías: gordo, flaco

La figura 5.1 muestra una población hipotética que tiene 18 individuos que son gordos, o flacos (no hay gente con peso normal en esta población) y una muestra representativa.

En la figura 5,1 podemos ver la población total y la muestra. Respecto de la población, podemos decir que: 5 de cada 9 personas son gordas. Esta relación se repite en la muestra representativa.



Figura 5.1 Muestra representativa de una población que sólo tiene una característica, el peso, con dos posibilidades: gordo y flaco.

En la vida real, es muy difícil que una muestra tenga proporciones idénticas a las poblacionales, pero **deberían ser muy parecidas** en todas las características que se puedan conocer.

Si se quiere realizar un estudio para averiguar a qué edad caminan los bebés de la Argentina, la muestra debería tener una distribución geográfica, por provincia o región, similar a la del último censo disponible de población. Si se considera que hay otros factores que pueden influir además de la región, por ejemplo el tipo de vivienda (casa o departamento), la muestra también tendrá que ser representativa de esos otros factores.

6. Muestreo

La forma en que se realiza la selección puede hacer la diferencia. Es más fácil obtener muestras malas que buenas.

No todo es tan simple como tomar sopa.

En la Sección 5.3 consideramos un estudio, realizado por un profesor universitario entre sus alumnos, para evaluar si los jóvenes que practican deportes consumen menos alcohol. Este es un ejemplo de participación voluntaria en un estudio, la muestra no es representativa de la población de interés.

Recordemos un ejemplo de la Sección 3.1. Interesaba conocer las opiniones respecto a si el correo electrónico debe responderse lo más rápido posible o no. Si la encuesta fue realizada vía el correo electrónico, las opiniones representan únicamente a los que tienen correo electrónico y les interesó responder la encuesta.

La próxima vez que se encuentre con un resultado de un estudio, averigüe qué composición tenía la muestra y pregúntese si la muestra representa a la población que interesa o a un subgrupo más restringido.

6.1 Muestreo aleatorio simple

Es bueno que la **muestra** se seleccione en forma **aleatoria;** esto significa que:

Cada uno de los individuos de la población tiene la misma oportunidad de ser seleccionado.

- Se utiliza algún mecanismo probabilístico para elegirlos.
- La gente no se selecciona a sí misma para participar.
- Nadie en la población es favorecido en el proceso de selección.

Muestra aleatoria simple: Una muestra aleatoria simple es la que se obtiene a partir de un mecanismo que le da a cada una de las unidades muestrales la misma probabilidad de ser elegida.

El muestreo aleatorio (el proceso por el cual se obtiene una muestra aleatoria) comienza con una lista de **unidades muestrales** de la que se extraerá la muestra. Esta lista se llama **marco muestral.** Idealmente, el marco muestral debería contener la lista de la totalidad de las unidades muestrales.

El **muestreo aleatorio simple** tiene dos propiedades que lo convierten en el procedimiento por excelencia de obtención de muestras.

- Todas las unidades tienen la misma oportunidad de ser elegidas (es insesgado).
- La elección de una unidad no influye sobre la elección de otra (independencia).

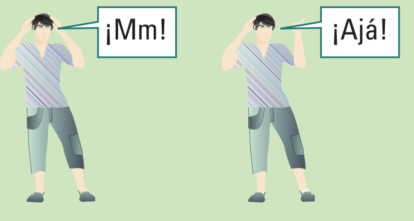
El Instituto Nacional de Estadísticas y Censos - INDEC - realiza periódicamente **censos** para registrar las características básicas sobre población y vivienda, actividad económica y agropecuaria de nuestro país. Las unidades relevadas en los censos proveen el **marco muestral** para las **encuestas** que realiza durante los períodos intercensales.

Se espera que el muestreo aleatorio provea muestras representativas de la población.

Mediante un censo se intenta registrar todas las unidades muestrales de la población para proveer el marco muestral. Si se trata de un censo de población, deberán localizarse todas las personas. Si se trata de un censo económico, se registrarán todos los locales comerciales y productivos. Una vez que se dispone del marco muestral se abre la oportunidad de seleccionar la muestra.

Por otra parte, es necesario aclarar que una unidad muestral puede contener muchos individuos. Una escuela, con sus alumnos, puede ser una unidad muestral. El objetivo del estudio pueden ser las escuelas (por ej. interesa conocer la superficie cubierta por alumno) o ser los alumnos (por ej. interesa conocer el rendimiento en educación física).





¿Cómo? ¿Una unidad muestral puede estar constituida por muchos individuos?

Volvamos al ejemplo de la encuesta sobre la preferencia del candidato a presidente del Club Grande de Fútbol. Utilicemos la lista actualizada de todos los socios como marco muestral con los números de socio para identificarlos. Si se decide que 1 de cada 6 socios entrarán en la muestra podemos arrojar un dado tantas veces como socios tenemos en la lista y si sale 1 el socio es seleccionado.

TABLA 6.1

Socio Número	Número aleatorio	Socio Número	Número aleatorio	Socio Número	Número aleatorio	Socio Número	Número aleatorio
1495	4	**1.501**	**1**	1.507	4	1.513	4
1496	8	1.502	6	1.508	4	1.514	7
1497	8	1.503	3	1.509	3	1.515	8
1498	7	1.504	7	1.510	8	1.516	8
1499	9	**1.505**	**1**	**1.511**	**1**	**1.517**	**1**
1500	5	1.506	7	1.512	7	1.518	3

Con este procedimiento, seleccionamos los socios no: 1.501, 1.505, 1.511 y 1.517 mediante un **muestreo aleatorio simple.**

También podríamos utilizar un programa de computadora para generar números entre 1 y 6 en forma aleatoria, sin necesidad de arrojar un dado.

Muestra aleatoria simple en dos pasos :
Paso 1. Se asigna una etiqueta numérica a cada individuo de la población.
Paso 2: Se utilizan números aleatorios para seleccionar las etiquetas al azar.

En la práctica, el primer paso del procedimiento es el más difícil. Esta dificultad da lugar a **muestreos alternativos** que **no** son **válidos** desde el punto de vista del análisis estadístico. Veremos algunos en la próxima sección.

□ 6.2 Muestras malas

Todos los días encontramos ejemplos de **muestras malas:**

- Cuando se pide a los oyentes de un programa de radio que voten por tal o cual cantante, llamando por teléfono o enviando un mensaje de correo electrónico, se trata de **muestras de respuesta voluntaria.** Las encuestas de opinión en las que se llama, o se escribe, por propia iniciativa son ejemplos de muestras de respuesta voluntaria, poco satisfactorias desde un punto de vista estadístico.

Cada vez que mire los resultados de un estudio, busque la frase "muestra aleatoria". Si la encuentra, hile más fino para averiguar cómo fue obtenida y si en realidad fue elegida en forma aleatoria.

- Otro tipo de muestra mala es la **muestra de conveniencia.** Si una pedagoga elige a sus propios alumnos, del último año de la escuela secundaria en la que trabaja, para evaluar un cambio en el método de enseñanza, los resultados no se podrán extender más allá de ese grupo.

□ 6.3 Sesgo

Alguna vez escuchamos **el sesgo es malo.** Pero, ¿qué es el sesgo? Es un favoritismo de alguna etapa del proceso de recolección de datos **beneficiando algunos resultados, perjudicando otros y desviando las conclusiones en direcciones equivocadas.**

> Cuando alguna etapa del proceso de recolección de datos está sesgada, **utilizar una muestra grande no corrige el error,** simplemente lo repite.

Los datos en un estudio pueden estar sesgados por muchos motivos. A continuación, veremos algunos de ellos.

6.3.1 Sesgo por elección de la muestra

6.3.1.1 Muestras por conveniencia

Exprimir las naranjas que se encuentran a la vista, en la parte de arriba del cajón, es un ejemplo de muestra de conveniencia. Las entrevistas en los centros comerciales (shopping) son otro ejemplo, porque los fabricantes y las agencias de publicidad suelen recolectar información respecto a los hábitos de compras de la población y el efecto de sus publicidades en grandes centros de compras. Obtener una muestra de esta manera es rápido y económico, pero la gente que contactan no es representativa de la mayoría de la población.

6.3.1.2 Muestras con sesgo personal

Por simpatía, gusto o interés, quien está realizando la encuesta puede preferir encuestar a cierto tipo de personas y no a otras. Por ejemplo, es posible que un encuestador joven tienda a buscar chicas bonitas para preguntarles.

6.3.1.3 Muestras de respuesta voluntaria

Surgen a partir de los individuos que se ofrecen voluntariamente a participar. Se trata, por ejemplo, de las que alimentan las votaciones organizadas por programas de radio, televisión o de algún sitio de Internet. No producen resultados que tengan algún significado en relación a la opinión de la población en general. Los participantes voluntarios, que por algún motivo decidieron participar, suelen tener opiniones más polarizadas.

6.3.2 Sesgo de respuesta



6.3.2.1 Debido a la presentación de las preguntas

Las diferentes palabras con las que se puede presentar una misma pregunta suele ser una fuente importante de sesgo en las respuestas.

En un curso de manejo organizado por un automóvil club se proyectó una película sobre un accidente de tránsito a dos grupos de alumnos. Ambos grupos eran similares respecto de la edad y el género. Al finalizar la proyección se preguntó:

- Al primer grupo: ¿a qué velocidad piensa que los dos autos chocaron? El promedio de las respuestas fue de 50,9 km/h.
- Al segundo grupo: ¿a qué velocidad piensa que los dos autos se colisionaron? El promedio de las respuestas fue de 65,9 km/h.

Ambos grupos vieron la misma película. El uso de la palabra **colisionaron** aumentó las estimaciones de la velocidad del accidente en **15 km/h,** esto es un aumento del 29,5 %

El sesgo debido a la forma en que se presenta una pregunta puede ser **intencional** o **no intencional.**

Las preguntas "¿No está usted harto de pagar impuestos para que todo siga igual de mal?" y "¿Le parece importante que se paguen impuestos para mejorar la educación, los servicios de salud y la seguridad?", que apuntan al pago de impuestos, seguramente tendrán resultados muy diferentes. Ambas preguntas conllevan un **sesgo intencional.**

Una encuesta dirigida a alumnos de 7mo. grado que pregunte: "¿Cuáles son las 5 personas grandes que le gustarían conocer personalmente?" tendrá diferentes lecturas. Algunos de los alumnos podrán interpretar que se trata de personas mayores de edad, otros que son altos, otros que se refiere a gordos o tal vez a grandes estrellas de cine, de rock, políticos o deportistas, generando un sesgo **no intencional.**

6.3.2.2 Para tratar de agradar

A la gente no le gusta mostrarse con ideas que no están bien vistas socialmente. Por ejemplo, cuando esté cara a cara con un encuestador o llenando un formulario no anónimo, un varón evitará una respuesta que parezca machista, o una mujer responderá tratando de ocultar algún prejuicio.

6.3.2.3 Por recuerdo

Si la pregunta está referida a un acontecimiento ocurrido algún tiempo atrás, la respuesta tendrá un sesgo por recuerdo. Por ejemplo, si se le pregunta a una madre a qué edad comenzaron a caminar sus hijos, la veracidad y precisión de la respuesta dependerá de las características personales de la madre.

6.3.2.4 Por no respuesta

Algunas veces las personas que han sido seleccionadas para una encuesta son muy difíciles de localizar o simplemente se niegan a responder. **Los individuos que no responden pueden ser muy diferentes de los que sí lo hacen.** Este tipo de sesgo se puede reducir sustituyendo a los se niegan a responder por otros individuos con las mismas características de los que no responden, pero suele ser difícil.

Cuando mire los resultados de una encuesta que le interesa especialmente, antes de sacar sus propias conclusiones averigüe qué se preguntó, cómo fueron redactadas las preguntas, si las respuestas fueron dadas en forma anónima o no y cuántos se negaron a responder.

Es más fácil obtener muestras malas que buenas.

6.3.2.5 Por subcubrimiento

Una encuesta telefónica ignora a todos los sujetos que no tienen teléfono. Una encuesta que realiza las entrevistas en hogares ignora a los que viven en la calle.

□ 6.4 Otros tipos de muestreos

6.4.1 Muestreo sistemático

Veamos un ejemplo de la utilidad de este método. Si nos interesa la opinión de las alumnas de una escuela respecto del aumento de las horas destinadas a la práctica de deportes, podríamos entrevistar a las alumnas a la salida y elegir una de cada diez (suponiendo que salgan de a una) hasta que hayan salido todas. De esta manera, si la escuela tiene 227 alumnas, la muestra tendrá 22 alumnas.

Muestreo sistemático: El muestreo comienza con una unidad elegida al azar y a partir de allí continúa cada k unidades. Si n es el tamaño muestral y N es el tamaño de la población entonces k es aproximadamente N/n.

Este tipo de muestreo permite evitar el sesgo personal y es más sencillo que el muestreo aleatorio. Es útil cuando la población está ordenada naturalmente (si no lo está, para utilizar este tipo de muestreo es necesario ordenarla, pero al ordenarla, se pierden las ventajas que tiene).

Por su simplicidad, se suele utilizar para **control de calidad** durante, o al finalizar, la fabricación de diversos productos.

En una producción continua de tubos de pasta dentífrica, se elige un tubo por hora y se lo analiza para verificar que cumple con las especificaciones.



Advertencia: Este muestreo no es adecuado cuando el período de la selección está relacionado con alguna característica que nos interesa evaluar.

Podría ocurrir que cada hora (una hora es el período de la selección) se produzca una leve caída de tensión que hace que los tubos de pasta dentífrica se llenen más o menos. No detectaríamos esa variación con el muestreo cada hora.

Al realizar un muestreo sistemático es importante estar alerta para identificar los factores que puedan estar invalidando los resultados.

6.4.2 Muestreo aleatorio estratificado

En un muestreo estratificado la población se divide en **grupos homogéneos** llamados estratos. Luego se realiza un muestreo aleatorio simple de unidades muestrales dentro de cada estrato.

Los estratos se eligen de acuerdo con los valores conocidos de algunas variables, de manera que haya **poca variabilidad dentro del estrato** (los valores de dichas variables para las unidades de un estrato particular difieren poco), pero que haya **mucha variabilidad entre estratos** (los valores de dichas variables para las unidades de distintos estratos difieren mucho).

Ejemplo 1:

La población de una ciudad podría estratificarse por

- **grupo de edad:** menos de 6 años, entre 6 y 12 años, entre 13 y 18 años y mayores de 18 años.
- **género:** femenino, masculino.

Así obtenemos 8 estratos, dentro de los cuales los individuos tienen 2 características similares: grupo de edad y género. Podríamos realizar un muestreo proporcional a la cantidad de individuos que tiene cada estrato, de manera que el tamaño de la muestra dentro de cada estrato dependa de la proporción de la población total que dicho estrato representa.

Ejemplo 2:

En una encuesta diseñada para conocer la situación de la industria en una provincia podrían utilizarse estratos por tamaño y actividad. Para cada actividad industrial podrían incluirse **todos** los locales industriales con 500 ó más obreros ocupados (**inclusión forzosa** - la muestra los contiene a todos), **la mitad** de los que tuvieran entre 499 y 200, **la cuarta parte** entre 199 a 50 y **1 de cada 20** para los de menos de 50. Tendríamos así 4 estratos:

- Estrato 1: Locales con 500 ó más obreros
- Estrato 2: Locales con 499-200 obreros
- Estrato 3: Locales con 199-50 obreros
- Estrato 4: Locales con 50-0 obreros

Si además se dividiera la actividad industrial en dos: 1) industria alimenticia, 2) industria no alimenticia, ¿cuántos estratos tendría la muestra? Tendría 8 estratos, dos por cada uno de los 4 estratos anteriores.

Tres pasos de un muestreo aleatorio estratificado:

- **Paso 1:** las unidades se agrupan en estratos. Los estratos se eligen teniendo en cuenta que estos grupos tienen un interés especial dentro de la población, o porque los individuos en el estrato se parecen mucho.
- **Paso 2:** se establece la proporción de unidades, o **fracción de muestreo,** que se incluirá para cada estrato
- **Paso 3:** dentro de cada estrato se realiza un muestreo aleatorio simple y la **proporción de individuos** que se incluye en la muestra es la establecida en el paso 2. La unión de las muestras de cada estrato constituye la muestra completa.

6.4.3 Muestreo por conglomerados

En un muestro por conglomerados la población se divide en **grupos heterogéneos** llamados **conglomerados.** Luego se realiza un muestreo aleatorio simple en el que las unidades muestrales son los conglomerados.

La idea del agrupamiento para un **muestreo aleatorio por conglomerados** (también llamados aglomerados) es opuesta a la del muestreo estratificado. Interesa que los individuos que componen cada grupo sean lo más heterogéneos posibles y se espera que cada conglomerado sea representativo de la población. Los **conglomerados** son las **unidades del muestreo,** pero las unidades de interés son los individuos dentro de los conglomerados. Se selecciona una muestra aleatoria de conglomerados, y **se observan todos los individuos dentro de cada conglomerado** ó se selecciona una muestra aleatoria simple dentro del conglomerado. Este tipo de muestreo puede tener mejor rendimiento costo-efectividad que un muestreo aleatorio simple, en especial si los costos de traslado son altos.

Ejemplo 1:

Una encuesta de viviendas. Se divide la ciudad en manzanas, se seleccionan las manzanas mediante un muestreo aleatorio simple y se visitan todas las casas de cada manzana seleccionada.

Ejemplo 2:

En un estudio interesa evaluar la capacidad de lectoescritura de alumnos de 7mo grado. Se seleccionarán al azar las escuelas y luego se realizará la prueba en todos los alumnos de 7mo. grado de las escuelas seleccionadas.

Tres pasos de un muestreo aleatorio por conglomerados:

- **Paso 1:** Los individuos se agrupan en conglomerados. Los conglomerados generalmente tienen una proximidad física, pero dentro de cada conglomerado las unidades son heterogéneas.
- **Paso 2:** Los conglomerados son las unidades muestrales. Se establece la proporción de unidades que se incluirá.
- **Paso 3:** Se realiza un muestreo aleatorio simple de conglomerados y se estudian todos los individuos de cada conglomerado seleccionado. El tamaño final de la muestra es la cantidad de individuos que componen todos los conglomerados seleccionados.

6.4.4 Muestreo multietápico

Un muestreo multietápico tiene dos o más pasos y, en cada uno de ellos se aplica cualquiera de los procedimientos de selección anteriores.

Ejemplo 1:

Una encuesta de viviendas. En **la primera etapa** se divide la ciudad en barrios, se toma una muestra aleatoria simple de barrios. En **la segunda etapa,** cada barrio seleccionado en la primera etapa se divide en manzanas, se seleccionan las manzanas mediante un muestreo aleatorio simple, y se visitan todas las casas de cada manzana seleccionada.

Ejemplo 2:

Estudio para evaluar la capacidad de lectoescritura de alumnos de 7mo. grado. En **la primera etapa** se seleccionan al azar las escuelas, y en **la segunda etapa** se selecciona dentro de cada escuela un cierto número de cursos de 7mo. grado. La prueba se realiza en todos los alumnos de 7mo. grado de los cursos seleccionados en la segunda etapa.

6.5 Actividades y ejercicios

1. ¿Cuál es la Población? ¿Cuál es la muestra?
   Para cada uno de los siguientes estudios indicar la población lo más detalladamente posible, es decir describir a los individuos que la componen. Si la información es insuficiente, completarla de la forma que se considere más adecuada. También indicar cuál es la muestra.

   - Una encuesta de opinión contacta a 1.243 adultos y les pregunta, ¿ha comprado un billete de lotería en los últimos 12 meses?
   - Durante la reunión anual del colegio de abogados, todos los presentes (2.500), llenaron una encuesta referida al tipo de seguro que prefería para su automóvil.
   - En 1968 se realizó en Holanda un test de inteligencia a todos los varones de 18 años que estaban realizando el Servicio Militar Obligatorio.
   - El INDEC lleva a cabo la Encuesta Permanente de Hogares (EPH) en la que se encuestan 25.000 hogares para captar información sobre la realidad económico-social de la República Argentina.

2. Voto secreto y obligatorio.

   - ¿Qué tipos de sesgos se pueden producir cuando una elección para presidente se realiza en forma voluntaria?
   - ¿Qué tipos de sesgos se pueden producir si el voto en la Comisión Directiva de un club o en la Cámara de Diputados no es secreto?

3. Se quiere realizar una encuesta entre los alumnos de una escuela secundaria, de 2.500 alumnos (500 alumnos por cada año, de 1ro. a 5to.), utilizando una muestra de tamaño 100. El propósito de la encuesta es determinar si a los/as alumno/as les interesa discutir el siguiente tema: "Debe reducirse la edad de imputabilidad penal para los menores de edad, que establece la ley nacional 22.278, a dieciséis años de edad; como respuesta al incremento en la cantidad de delitos graves cometidos por jóvenes y adolescentes".

4. Indicar cuál es el tipo de muestreo realizado en cada caso.

   - Cada alumno escribe su nombre en un papel, lo pone en una bolsa y el director elige 100 papeles.
   - A cada alumno se le asigna un número entre 1 y 2.500 y se seleccionan generando 100 números al azar de cuatro dígitos utilizando algún programa de computación.
   - Para cada año se asigna a cada alumno un número entre 1 y 500, y se elige 1 de cada 25 alumnos.
   - Se eligen al azar una división de cada uno de los años y se seleccionan 20 alumnos de cada división.
   - Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años

- Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años. Se seleccionan en forma separada los varones y las mujeres de acuerdo con la proporción de mujeres y varones que tiene la escuela.

5. En un programa de radio se invitó a las/los oyentes a contestar la siguiente pregunta: "¿Si pudiera volver el tiempo atrás volvería a tener hijos?" De más de 10.000 respuestas el 70% dijo no. ¿Qué muestra esto?

   Elegir, entre las cinco siguientes, la respuesta que mejor responde a esta última pregunta.

   a. La encuesta no dice nada porque arrastra el sesgo por respuesta voluntaria.
   b. No se puede decir nada sin saber las características de los oyentes.
   c. Para sacar una conclusión, es necesario separar las respuestas entre hombres y mujeres.
   d. Hubiese tenido más sentido tomar una muestra aleatoria de las 10.000 respuestas para sacar conclusiones.
   e. Es una muestra legítima elegida al azar entre todos los que escuchan ese programa y tiene un tamaño suficiente como para concluir que la mayoría de los oyentes lo pensarían dos veces antes de tener más hijos.

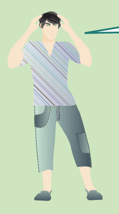
6. Indicar cuál o cuáles de las siguientes afirmaciones son válidas.

   a. Las respuestas que se obtienen al utilizar un cuestionario expresado en términos no neutrales tendrán "sesgo por respuestas".
   b. Las encuestas de respuesta voluntaria subestiman a la gente con opiniones muy firmes.
   c. Las encuestas de respuesta voluntaria generalmente sobre representan las respuestas negativas.
   d. En general, es posible reducir el sesgo tomando muestras muy grandes, cuanto más grande es el tamaño de la muestra mejor.
   e. El tamaño de la muestra no tiene nada que ver con el sesgo.
   f. Los resultados que se obtienen de un censo son siempre más precisos que los que se obtienen de una muestra, sin que importe cuán cuidadoso haya sido el diseño y su aplicación.

7. Datos - variables

Los datos numéricos son valores de variables numéricas.
Los datos categóricos son valores de variables categóricas.

Las **variables** son **características** que pueden tomar valores diferentes de una unidad a otra, como la edad de las personas, la cantidad de habitantes de cada ciudad, la duración o el consumo de una lamparita.



¿Datos y variables? ¿Son o no son lo mismo?

¿Entonces que son los datos?

Los **datos** son los **valores** observados de las variables.

Para ilustrar los conceptos, consideremos la siguiente tabla. Muestra una parte de la libreta donde la maestra registra datos de sus alumnos.

Alumno	Lengua	Matemática	Ciencias Naturales	Participación	Certificado de Vacunas
Cortez María	8,25	6,12	9,51	Buena	Si
García Lobos, Federico	6,59	9,06	8,47	Regular	Si
Gordon, Susana	9,07	7,39	9,72	Buena	Si
Medignone, Horacio	7,55	6,42	8,64	Mala	No
Vázquez, Florencia	6,25	9,63	7,59	Buena	Si

Las unidades son los alumnos del grado, identificados mediante la variable "Alumno" cuyos valores son el nombre y apellido de cada uno de ellos (primera columna de la tabla). Las cinco columnas restantes contienen el **nombre** y los valores de las demás **variables.**

Los nombres encabezan las columnas: Lengua, Matemática, Ciencias Naturales, Participación, Certificado de Vacunas, y en el cuerpo de la tabla (filas a continuación) aparecen los **valores** de cada una de ellas.

Nombres de las variables
Valores observados de las variables (datos)

Las variables tienen un **nombre** y un **valor** para cada individuo de la población.

Los **datos** son los **valores** observados -medidos- **de las variables** para los individuos de una muestra.

Los datos solos dicen muy poco, si no sabemos a qué variables corresponden.

7.1 Variables numéricas y variables categóricas

Los **datos numéricos** son valores de variables numéricas. Los **datos categóricos** son valores de variables categóricas.

En el ejemplo de la libreta de anotaciones de una maestra, las columnas 2, 3 y 4 dan el promedio de notas en cada una de las asignaturas, se trata de **variables numéricas.** La primera, muestra el nombre y apellido de cada alumno; la quinta, el grado de participación en clase registrado en 3 categorías, y la sexta, si la/el alumna/o presentó o no presentó su certificado de vacunas. Todas ellas son **variables categóricas.**

La estadística trata con números, pero **no todas las variables son numéricas.** En este ejemplo, la primera y las dos últimas son **categóricas.** Para resumir los valores de este tipo de variables utilizamos **cantidades y porcentajes.** Por ejemplo, podemos calcular la cantidad de alumnos que se llaman "Juan", o que entregaron el certificado de vacunas, o el porcentaje de alumnas/os que tienen una participación "Buena".

La mayoría de las variables (y por consecuencia también de los datos) se pueden clasificar en **numéricas** y **categóricas.** También se los denominan **cuantitativos** y **cualitativos** respectivamente.

Para analizar variables categóricas se utilizan **cantidades, proporciones** y **porcentajes.**

Ejemplo:

En el censo de población de la República Argentina del año 2001, una de las preguntas fue: ¿Cuál es el grado de educación de las personas con 15 años y más?
La tabla 7.1 responde a esa pregunta. Su título permite ver, inmediatamente, de qué se tratan los datos. Se consigna el año porque estos datos cambian con el tiempo.

Al pie figura, la fuente de los datos: el INDEC. En la primera columna de la tabla se presentan los nombres de las categorías de la variable "Nivel de Educación"; en la segunda y tercera su distribución. En la segunda columna, la distribución se expresa en cantidades, con el encabezamiento indicando "Cantidad de personas". En la tercera columna, la distribución se expresa en porcentajes como también lo muestra su encabezamiento. Suele ser más sencillo pensar en porcentajes. Es más fácil decir el 48,9% tiene estudios primarios completos, que decir que 12.720.081 personas tienen estudios primarios completos.

Distribución del Nivel de educación de la población de 15 años y más. 2001 TABLA 7.1

Nivel de Educación	Cantidad de personas	Porcentaje
Sin instrucción (1)	962.460	3,7
Primario incompleto	3.693.766	14,2
Primario completo	12.720.081	48,9
Secundario completo	6.373.046	24,5
Terciario completo	2.263.082	8,7
Total	26.012.435	100

(1) incluye nunca asistió, jardín e inicial.

Fuente: *INDEC. Dirección Nacional de Estadísticas Sociales y de Población. Dirección de Estadísticas Sectoriales en base a procesamientos especiales del Censo Nacional de Población, Hogares y Viviendas 2001.*

Distribución de una variable: La distribución de una variable nos dice cuáles son sus posibles valores y con qué frecuencia aparecen.

La tabla 7.1 muestra la distribución de la **variable categórica "Nivel de educación"**, máximo nivel de educación alcanzado por las personas de 15 años o más. Tiene 5 categorías: "Sin instrucción", "Primario incompleto", "Primario completo", "Secundario completo" y "Terciario completo". La columna encabezada por "Cantidad de personas" muestra **la frecuencia de cada una de las 5 categorías,** esto es, la **cantidad de personas** que pertenecen a esa categoría. Se trata de **frecuencias absolutas. La suma de las frecuencias** da como resultado la **cantidad total de datos,** 26.012.435, es la cantidad de personas de 15 años ó más en el año 2001.

La frecuencia relativa es el cociente entre la frecuencia absoluta y la cantidad total de datos. Su suma es 1. Cuando las frecuencias relativas están expresadas en porcentaje, la suma es 100, como vemos en la tercera columna de la tabla 7.1.

7.1.1 Gráficos para datos categóricos

7.1.1.1 Gráficos circulares

Utilizaremos un gráfico circular, también llamado gráfico de torta, para visualizar la distribución de la variable "nivel de educación" (tabla 7.1). Podremos visualizar los porcentajes de personas que pertenecen a cada una de las 5 categorías.



Gráfico circular: Se utiliza para representar la distribución de los valores de una variable categórica. El círculo representa el total de los datos. Cada sector dentro del círculo representa una categoría con el ángulo proporcional a su tamaño (cantidad o porcentaje que pertenece a dicha categoría).

Para realizar un gráfico circular, primero se dibuja un círculo. Los 360º representan el total, en este caso todas las personas de 15 años o más de la República Argentina en el 2001. Cada sector dentro del círculo representa una categoría con el ángulo proporcional a su tamaño (cantidad o porcentaje). El sector correspondiente a la categoría "Secundario completo" tendrá un ángulo de 0,245 x 360=88,2 grados.

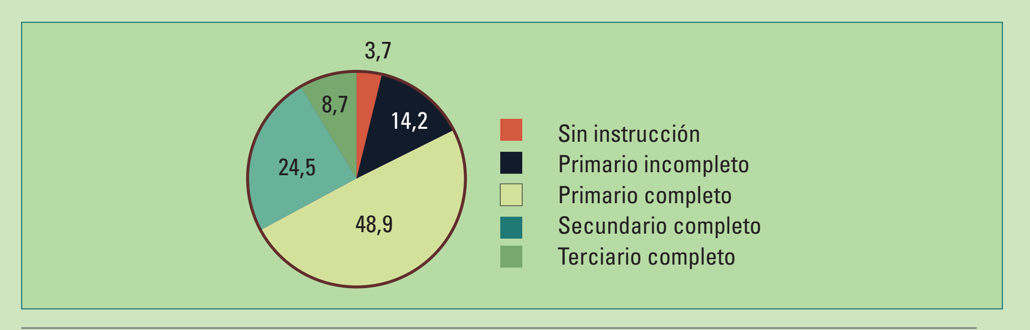


Figura 7.1. *Gráfico circular de la distribución del nivel de educación de las personas de 15 años y más de la República Argentina. Año 2001*

Los gráficos circulares permiten visualizar cómo las partes forman el total, aunque es más difícil comparar ángulos que longitudes. Estos gráficos no son buenos para comparar con precisión los tamaños de las diferentes partes, para eso lo gráficos de barras son mejores.

Los gráficos circulares muestran sectores de área proporcional al porcentaje del total correspondiente a cada grupo o categoría, pero generalmente no muestran la cantidad total en cada grupo, en términos de unidades originales (pesos, número de personas, etc.). Este enfoque se traduce en una pérdida de información.

Para ilustrar esa situación consideremos los datos proporcionados por la Lotería de la Provincia de Buenos Aires en Junio de 2008 http://www.loteria.gba.gov.ar/ sobre como reparte sus ganancias entre diferentes organismos de la provincia.

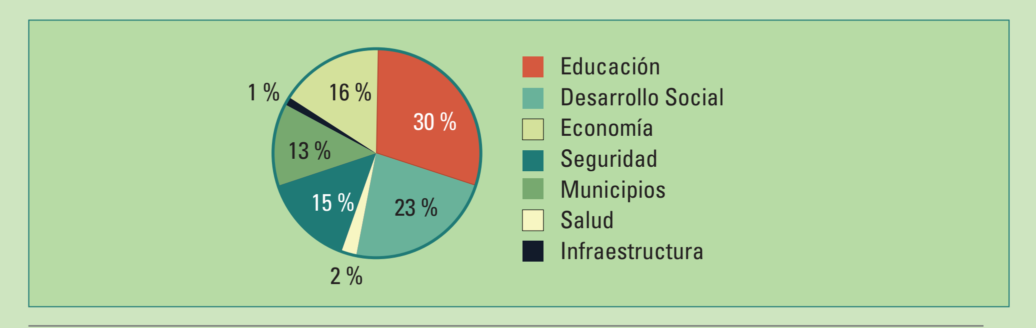


Figura 7.2. *Gráfico circular de la distribución las ganancias de la Lotería de la Provincia de Buenos Aires junio de 2008*

Vemos los porcentajes destinados a los diferentes organismos. Se destinó más del 50% entre Educación y Desarrollo Social. Pero, ¿cuánto fue realmente, en pesos? Veamos esa información en la tabla siguiente.

Siempre se puede pasar de cantidades a porcentajes. En la página de la Lotería de la provincia de Buenos Aires aparecen las cantidades totales y las destinadas a educación por mes, para el período enero-julio de 2008, pero aunque no están los porcentajes podemos calcularlos:

Año 2008	Educación	Total mensual	Porcentaje
Enero	37.307.382	143.225.097	26%
Febrero	45.541.083	164.313.370	28%
Marzo	34.872.907	130.834.379	27%
Abril	32.646.300	116.425.710	28%
Mayo	25.241.707	96.293.288	26%
Junio	35.416.187	117.960.104	30%
Julio	45.553.614	139.475.636	33%

No se puede ir de los porcentajes a los valores originales sin el conocimiento del total. Esta falta de información puede ser un verdadero problema, por ejemplo, cuando los gráficos muestran los resultados de una encuesta de opinión. Para evaluar el margen de error del porcentaje de personas que respondieron a la pregunta de una manera determinada es necesario saber cuántas personas respondieron la encuesta.

Distribución de las ganancias de la Lotería de la provincia de Buenos Aires de junio de 2008 por organismo. TABLA 7.2

Organismo	Junio 2008
Educación	35.416.187
Desarrollo Social	27.370.667
Salud	2.843.829
Seguridad	17.224.945
Municipios	15.141.832
Infraestructura	1.413.519
Economía	18.549.125
Total	**117.960.104**

7.1.1.2 Gráficos de barras

Los gráficos de barras se utilizan para representar la distribución de los valores de una variable categórica.

Las categorías se representan en el eje horizontal y la cantidad, o el porcentaje, de datos en el eje vertical. **La altura de las barras** sobre cada categoría representa la cantidad de datos de cada una de ellas. Tal como ocurre con los gráficos circulares, divide a los datos en grupos correspondientes a las categorías y muestra cuántos, o qué porcentaje de individuos pertenecen a cada categoría. Mientras que los gráficos circulares utilizan fundamentalmente porcentajes para indicar el tamaño de cada clase, los gráficos de barras utilizan tanto cantidades como porcentajes.

La figura 7.3 muestra un gráfico de barras de la distribución de los valores de la variable “Nivel de Educación”. La altura de cada barra representa los porcentajes de las personas de más de 15 años con nivel de educación mostrado en su base. La barra sobre la categoría “Primario Completo” es la más alta, es la categoría con la mayor cantidad de personas. Podemos comparar categorías: vemos que son más los individuos que tienen el secundario completo, que aquellos que no completaron su educación primaria.

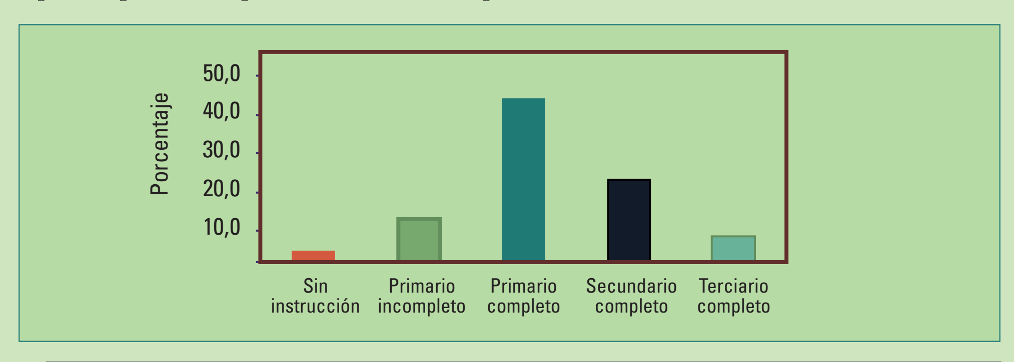


Figura 7.3. *Gráfico de barras de la distribución de la población de 15 años y más de la República Argentina, según máximo nivel educativo. Año 2001*

El gráfico de barras tiene un interés adicional cuando las categorías tienen un orden natural como ocurre en este caso. Vemos que la categoría central “nivel primario completo” es la más poblada y que la caída es más abrupta hacia las categorías correspondientes a menores niveles de educación que hacia los mayores.

Tanto en los gráficos de barras como en los gráficos circulares, los porcentajes de las categorías tienen que sumar 100%:

$$3{,}7\ \% + 14{,}2\% + 48{,}9\% + 24{,}5\% + 8{,}7\% = 100\%$$

7.1.2 Dos variables categóricas

Retomando el tema del nivel de educación, el INDEC incluye los totales y los porcentajes por nivel de educación y género en la presentación de la información de la distribución de la población de 15 años y más de la República Argentina. La tabla nos muestra cómo se distribuyen en forma conjunta dos variables categóricas, nivel de educación y género.

Podemos calcular las cantidades de todas las casillas que nos interesen.

DISTRIBUCIÓN DE LA POBLACIÓN DE 15 AÑOS O MÁS SEGÚN NIVEL DE EDUCACIÓN DE Y GÉNERO. AÑO 2001 TABLA 7.3

Nivel de educación	Total	Total	Género	
			Varón	Mujer
		26.012.435	12.456.479	13.555.956
	Sin instrucción (1)	**3,7%**	3,5%	3,9%
	Primario incompleto	**14,2%**	14,3%	14,1%
	Primario completo	**48,9%**	51,5%	46,5%
	Secundario completo	**24,5%**	23,7%	25,2%
	Terciario completo	**8,7%**	7,0%	10,3%

(1) incluye nunca asistió, jardín e inicial.

Fuente: *INDEC. Dirección Nacional de Estadísticas Sociales y de Población. Dirección de Estadísticas Sectoriales en base a procesamientos especiales del Censo Nacional de Población, Hogares y Viviendas 2001*

A menudo los **gráficos de barras** se utilizan para **comparar dos grupos,** dividiendo la barra de cada categoría en dos y mostrándolas una al lado de la otra.

Un gráfico de barras conjunto nos permite comparar las distribuciones de la variable "Nivel de Educación" en varones y mujeres.

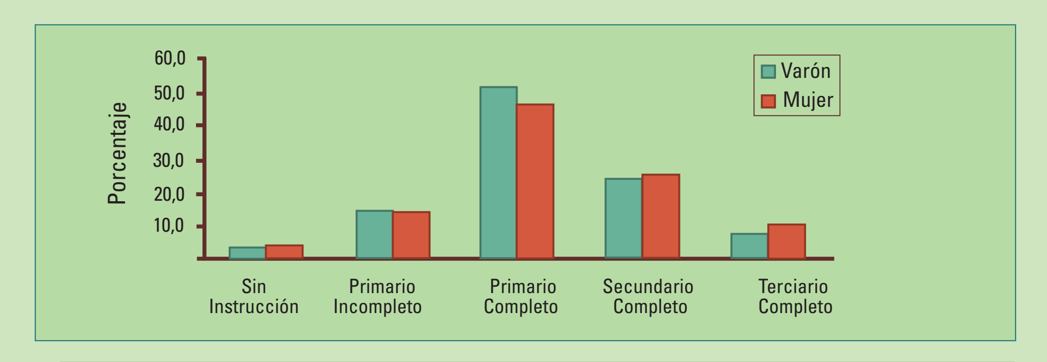


Figura 7.4. *Porcentaje de personas con más de 15 años de acuerdo al nivel de educación y género. Datos tabla 7.3.*

Vemos que en el nivel primario hay más varones que mujeres, pero en el secundario y terciario la relación se invierte, aunque todas las diferencias son pequeñas.

7.2 Actividades y ejercicios

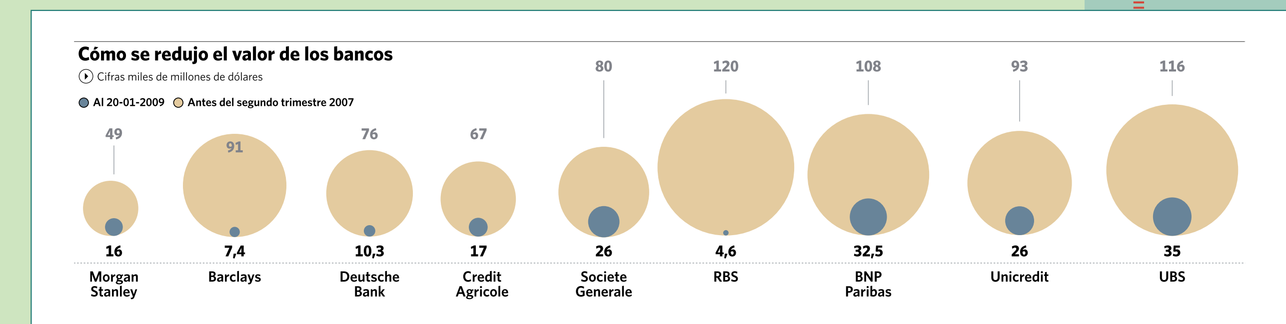
1. Un pictograma es un gráfico de barras que se reemplaza por figuras. Las figuras representan las cantidades o los porcentajes. En forma intencional o no intencional, muchas veces los gráficos exageran las relaciones entre las categorías.

    a. Se utilizó el siguiente pictograma para ilustrar una reducción cercana al 50% de los abandonos de mascotas en la vía pública de una ciudad después de una campaña oficial de concientización. Para reflejar esa reducción sin distorsionar la figura, el artista redujo tanto el alto como el ancho en un 50%:



Explique por qué la sensación visual de la reducción es bastante mayor que 50%. ¿Cómo debería haber sido la reducción de la figura para reflejarla en forma adecuada?

    b. Un artículo referido a las consecuencias de la crisis financiera de Estados Unidos en 2008 ilustra la reducción de los valores de los bancos mediante el siguiente pictograma. El valor del banco se calcula multiplicando la cantidad total de acciones por su cotización en la Bolsa de Nueva York.



Fuente: *Diario Clarín, 22 de Febrero 2009*

Indique si el pictograma muestra en forma correcta la reducción. Observe que los diámetros de los círculos son proporcionales a los valores.

2. Los siguientes datos son parte de los resultados del primer censo general de la Provincia de Santa Fe (1887). http://www.digitalmicrofilm.com.ar/censos/estadisticas.php

Localización de la vivienda		**Nacionalidades**				**Alfabetización**	
Urbana	90.764	Argentina	92.170	Inglaterra	753	Sí sabe escribir	62.608
Rural	116.712	Italia	46.268	Paraguay	673	No sabe escribir	87.042
Fluvial	2.250	Suiza	5.232	Chile	211		
Otros	382	Francia	2.944	Brasil	192		
		España	2.397	Bélgica	142		
		Alemania	2.070	Portugal	76		
		Austria	1.131	Estados Unidos	74		
		Uruguay	903				

Obtenga un diagrama de barras y un gráfico circular para distribución de los habitantes de la provincia de Santa Fe en 1887 de acuerdo con cada una de las siguientes tres variables categóricas: 1) Alfabetización, 2) Nacionalidades y 3) Localización de la vivienda.

3. Utilice el gráfico que considere adecuado para representar los datos de la tabla siguiente.

PRODUCTO BRUTO NOMINAL EN DÓLARES PER CÁPITA PARA 10 PAÍSES DE AMÉRICA DEL SUR, DURANTE 2008 SEGÚN EL FMI

Argentina	8.522	Ecuador	3.927
Bolivia	1.889	Paraguay	2.658
Brasil	8.676	Perú	4.610
Chile	10.814	Uruguay	8.860
Colombia	5.174	Venezuela	11.828

Fuente: *http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita*

8. Origen de los datos

¿Cuál es el tipo de estudio adecuado para responder una pregunta en particular?

8.1 Censos, encuestas, estudios observacionales y experimentales

Cuando un estudio se basa en consultar toda una población se denomina censo.

> **Censo:** El objetivo de un censo es obtener un registro de todos los miembros de una población en la forma más completa posible. Se relevan las variables principales que permitan la elaboración del marco muestral para futuras encuestas.

Por ejemplo, en un censo de población y vivienda se intenta contactar a todos los habitantes para obtener información respecto de edad, estado civil, género, ocupación y años de escolaridad. Algunas veces, también se recoge información sobre cuestiones relacionadas con las condiciones de vida de la población: en qué tipo de casa viven, si tienen acceso a servicios de salud y a educación, si tienen provisión de agua potable y sistema sanitario, etc. En la República Argentina, desde 1968, el INDEC es el encargado de conducir un censo de población cada 10 años. También realiza el Censo Nacional Agropecuario y el Censo Nacional Económico, con una periodicidad similar.

Un **estudio** puede ser **muestral,** cuando sus resultados están basados en información obtenida a partir una muestra. Este es el caso de la mayoría de los estudios de mercado, encuestas de opinión, evaluación de drogas, etc. En todos estos casos, el objetivo es obtener conclusiones respecto a la población de la que se obtuvo la muestra.

Para muchas investigaciones, provenientes tanto del ámbito privado como público, **no es razonable realizar un censo** por el tiempo y costo que involucra. Más aun, los intentos por recolectar datos completos de una población llevan muchas veces a información de menor calidad.

Un **estudio muestral bien diseñado,** y cuidadosamente conducido, es por lejos superior a un **estudio poblacional** (censo) con **mal diseño** o con escasos recursos. **Si una pregunta está redactada en forma confusa, sus respuestas pueden no tener ningún significado, aunque haya respondido toda la población.**

En una **encuesta** el interés está en **obtener información sobre toda la población estudiando una parte** de ella, es decir una **muestra.** Se intenta recoger información sin perturbar o modificar a la población para no afectar la calidad de los resultados.

Existen numerosos procedimientos para recoger la información a través de un muestreo. Gran parte de la información que nos llega surge de encuestas.

Los estudios pueden ser **observacionales** o **experimentales.**

Todos los **estudios observacionales** comparten un principio: **"sólo mire".** Si le interesa estudiar los hábitos de los pájaros de un bosque y para ganarse su confianza les ofrece migas de pan estará modificando su comportamiento, entonces ya no será un estudio observacional puro.

> Las encuestas son estudios observacionales, pero no todos los estudios observacionales son encuestas.

Por el contrario, en un **estudio experimental** se quiere modificar un comportamiento; no sólo observar a los individuos o realizar preguntas sin perturbar. **Se impone en forma deliberada una modificación** de las condiciones para observar las modificaciones generadas.

Por ejemplo, mediante estudios experimentales se podrán responder a las preguntas: ¿Bajaron de peso los alumnos que fueron obligados a duplicar sus horas de actividad física?, ¿Se redujo el índice de mortalidad infantil al habilitar instalaciones de agua potable en una población aislada?

□ 8.2 ¿Pueden estar mal los datos?

Permanentemente, en la televisión, diarios, blogs de Internet, vemos resultados del tipo:

- El 70 % de los que tienen entre 16 y 19 años piensan que bajar música de la red es lo mismo que comprar un CD usado o grabar música prestada de una amiga.
- Para los adolescentes, fumar ya comienza a ser mal visto.
- La depresión es causante de partos prematuros.
- El 49 % de los argentinos tiene sobrepeso.
- En el año 2008 el sueldo de un CEO, el máximo ejecutivo financiero de una corporación (por sus siglas en inglés Chief Executive Officer), era 23 veces más alto que el de un operario, mientras que en el año 1999 lo era 34 veces.

Los buenos datos son el resultado de un enfoque inteligente y un gran esfuerzo. Los datos malos resultan de la falta de cuidado, poco entendimiento del problema o incluso de la intención de producir resultados deliberadamente erróneos. Cuando escuchamos resultados impactantes como los anteriores, lo primero que tenemos que preguntarnos es: ¿con qué calidad de datos se obtuvieron?

☐ 8.3 Aspectos éticos

Cuando se recolectan datos de personas, tanto en estudios observacionales como experimentales, surgen complejos aspectos éticos. La situación más complicada se presenta en estudios experimentales en los que se impone un tratamiento a las personas. Los llamados ensayos clínicos son unos de los principales ejemplos. Un ensayo en el que se estudia un medicamento nuevo puede producir, por ejemplo, tanto un daño como un beneficio a los sujetos participantes.

A continuación, describiremos estándares básicos que se deben cumplir en la realización de un estudio que toma datos de personas, ya sea observacional o experimental:

- La organización que lleva adelante el estudio tiene que tener una junta que revise por adelantado los estudios planificados, para proteger a los sujetos participantes de un posible daño. Los hospitales suelen tener un **comité de ética** que se encarga de realizar ese control.
- Antes del inicio, todos los individuos que participan en el estudio tienen que dar un **consentimiento informado.** Tienen que ser informados, con anterioridad a la realización del estudio, sobre la naturaleza del mismo y el riesgo que ocurra algún daño.
    - En una encuesta (estudio observacional) no hay daño físico posible, se debe informar qué **tipo de preguntas** se realizarán y cuánto **tiempo** ocupará responderlas.
    - En los estudios experimentales los sujetos deben recibir la información sobre **la naturaleza y el objetivo del estudio y una descripción de los posibles riesgos.** Luego deben expresar su consentimiento, generalmente por escrito.

El organismo regulador de los ensayos clínicos de la Argentina es el ANMAT (Administración Nacional de Medicamentos y Tecnología Médica, www.anmat.gov.ar)

- Todos los datos individuales se deben guardar en forma confidencial. Solamente se pueden dar a conocer resultados resumidos para grupos de individuos.

☐ 8.4 ¿Cómo elegir un tipo de estudio?

¿Cuál es el tipo de estudio adecuado para responder a una pregunta en particular? Por ejemplo, si interesa conocer la opinión de ciertas personas, describir sus estilos de vida y preferencias o describir variables demográficas como nacimientos, muertes o migraciones, es adecuado realizar encuestas, sondeos y otros estudios observacionales. En cambio, si interesa determinar la causa de un resultado o comportamiento (es decir, una razón por la cual sucedió algo), un experimento es mucho mejor. Si no es posible (porque resulta inmoral, demasiado caro, o inviable), la realización de gran cantidad de estudios observacionales - analizando muchos factores diferentes - es la segunda mejor alternativa.

Veremos esto más adelante con mayor profundidad.

8.5 Actividades y ejercicios

1. Indicar y explicar cuál es el tipo de estudio más adecuado para responder a cada una de las siguientes preguntas:

   - ¿Están contentos los alumnos con el nuevo sistema de promoción?
   - ¿El ausentismo de los alumnos es menor en verano que en invierno?
   - ¿El rendimiento de los alumnos en un examen es mejor si durante el mismo escuchan música de Vivaldi, en bajo volumen, en comparación con no escuchar nada?

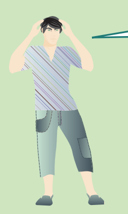
2. Presentar ejemplos de preguntas sobre los estudiantes de una escuela respecto a comportamiento, gustos y opiniones que podrían responderse con cada uno de los siguientes estudios:

   - Una encuesta
   - Un estudio observacional que no sea una encuesta
   - Un experimento

3. Una educadora divide al azar un grupo de niños y niñas de preescolar en dos grupos con iguales capacidades iniciales (para ello les toma una prueba). En un grupo utiliza canciones para enseñarles a contar, y en el otro el método tradicional. ¿Es esto un experimento? Explicar porqué sí o porqué no.

9. "Estadísticos" y "parámetros"

Cuando un estadístico se calcula en base a los datos de toda la población, ese resultado se denomina parámetro.



¿Parámetros? ¿Estadísticos? ¿Estimaciones?

Aunque la definición parezca nueva, ya nos hemos encontrado con **parámetros** y sus **estimaciones.**

En el ejemplo de las elecciones para presidente del Club Grande de Fútbol (sección 4.2), la verdadera **proporción** de **todos los socios** que están a favor del primer candidato es un **parámetro** que indicamos con la letra p. Describe a la **población de 58.210 socios del club.** Lo llamamos p por proporción, pero no lo conocemos.

La proporción que se obtiene a partir de una muestra es un **estadístico,** lo llamamos $\hat{p}$ (se lee p sombrero).

La investigadora finalmente obtuvo las respuestas de 538 socios, con 274 a favor del primer candidato. La **estimación del parámetro** es:

$$\hat{p} = \frac{275}{538} = 0{,}51$$

El 51% de los socios de la muestra está a favor del primer candidato, lo sabemos porque la investigadora se los preguntó. No sabemos cuál es el porcentaje real de todos los socios que lo apoyan, pero **estimamos** que alrededor de un 51% lo hace.

Consideremos nuevamente la población de todos los socios de un club, pero esta vez observemos su **edad.** El promedio de sus edades es un **parámetro,** lo llamamos **media** de la **variable edad.** Pero si seleccionamos una **muestra** de socios y calculamos el promedio de sus edades obtenemos una **media muestral.** La **media muestral** (capítulo 18) es un estadístico, cuyo valor depende de la **muestra elegida;** se parecerá a la media poblacional (el parámetro) pero en general no será igual.

La media poblacional generalmente se indica por la letra griega mu, **μ.**

Parámetros y estadísticos: Cuando el conjunto de datos proviene de la población completa, el valor del estadístico es un **parámetro.** Un **parámetro** es un número que describe la **población,** pero en la práctica casi nunca sabremos cuál es ese número porque no podemos conocer perfectamente a toda la población.

Cuando el conjunto de datos proviene de una muestra, el número obtenido es el **estadístico** que se utiliza como **una estimación del parámetro.**

La diferencia entre el parámetro y el estadístico es el **error de estimación.**

9.1 Actividades y ejercicios

En cada uno de los siguientes ejercicios

a) Indicar cuál es la unidad muestral, la variable, el estadístico, la población y, cuando corresponda, identificar el tamaño de la muestra.
b) Si el valor en negrita es un parámetro o el valor de un estadístico.

1. Un lote de arandelas tiene un diámetro promedio de **1,908** cm. Este valor se encuentra dentro de las especificaciones de aceptación del lote por parte del comprador. Un inspector selecciona 100 arandelas y obtiene un promedio de **1,915** cm de diámetro. Este valor se encuentra fuera de los especificados límites, por lo tanto el lote es rechazado erróneamente.

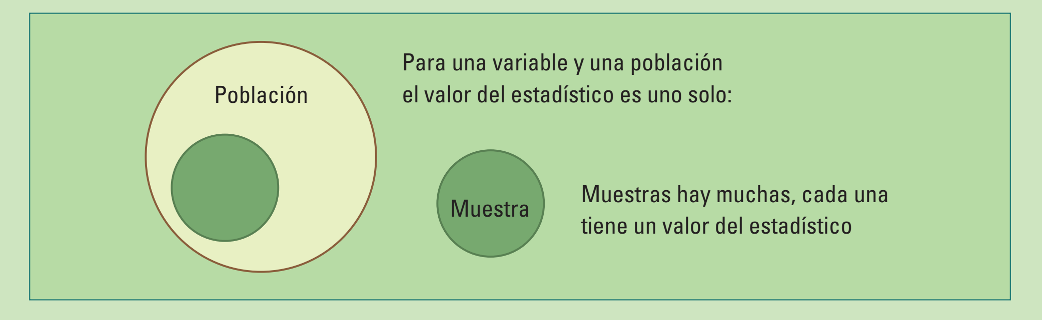
2. En un estudio reciente se entrevistaron 213 familias y la mayoría de las madres estaba al tanto de que los resfríos eran producidos por virus. Pero solamente el **40%** sabía que un antibiótico no puede curar un resfrío, y una de cada 5 creía, en forma equivocada, que un antibiótico lo podía prevenir.

3. En el año 2001 el **50%** de los hogares de la Argentina tenían heladera con freezer, de acuerdo con los valores censales del Anuario Estadístico de la República Argentina de 2006.

4. En el año 2009 el precio promedio de 8 autos modelo 2002 era de **$21.880.**

10. Variabilidad entre muestra y muestra

Siguiendo con el ejemplo del Club Grande de Fútbol, si la encuestadora obtuviera una segunda muestra aleatoria de 538 socios, la nueva muestra estaría compuesta por otros socios (alguno podría coincidir pero muchos serían diferentes). Es casi seguro que no habría exactamente 275 respuestas a favor del candidato 1 como ocurrió con la primera muestra. **Esto significa que el valor del $\hat{p}$ estadístico varía de muestra a muestra:** podría ocurrir que una muestra encuentre un 51% de socios a favor del candidato 1 mientras que una segunda sólo encuentre 37%.



La **primera ventaja** de las muestras aleatorias es que eliminan el **sesgo** del procedimiento de selección de una muestra. Aún así, suele no coincidir el resultado con el verdadero valor, debido a la **variabilidad** que resulta de la selección al azar. Este tipo de variabilidad es llamada **variabilidad muestral.**

Que la **variabilidad muestral** sea muy grande, significa que **el valor del estadístico cambia mucho entre muestra y muestra.** Por lo tanto no podemos creerle al resultado que obtenemos con una muestra en particular. Pero estamos salvados por una **segunda ventaja** que tienen las muestras aleatorias: la variación entre muestra y muestra (de un mismo tamaño) seguirá un patrón predecible. Este patrón predecible muestra que:

Los resultados de muestras de mayor tamaño son menos variables que los resultados de muestras más chicas.

□ 10.1 Muchas muestras

Para ver cuánto le podemos creer al resultado de una muestra debemos preguntarnos ¿qué pasaría si tomásemos muchas muestras de la misma población?

Probemos y veamos en el ejemplo de las elecciones del Club Grande de Fútbol. Supongamos que en realidad (esto no lo sabemos) la mitad de los socios del club (29.105) está a favor y la mitad en contra. Es decir, **la verdadera proporción** (parámetro) de socios que está a favor de uno de los dos candidatos es $p = 0{,}5$

Exactamente el 50% de los socios está a favor del candidato 1.

¿Qué pasaría si utilizáramos una muestra de tamaño 35? Es un tamaño bastante chico para estimar el valor desconocido p de la verdadera proporción de socios a favor del candidato 1.

La figura 10.1 ilustra el resultado de elegir **1.000 muestras diferentes** de tamaño 35 y hallar el valor de $\hat{p}$ para cada una de ellas.

En la primera, de las 1000 muestras, sólo 12 de las 35 personas prefirieron al candidato 1 resultando la proporción $\hat{p} = \frac{12}{35}$

$= 0{,}34$

En la segunda muestra 20 de las 35 personas prefirieron al candidato 1, resultando una estimación de p: $\hat{p} = \frac{20}{35}$ implica que: $\hat{p} = 0{,}57$ para la segunda muestra.

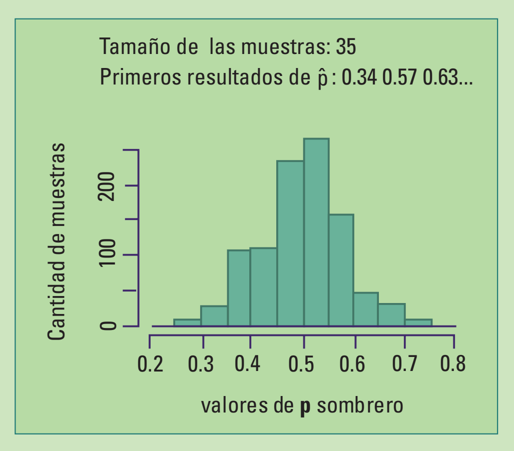


Figura 10.1.

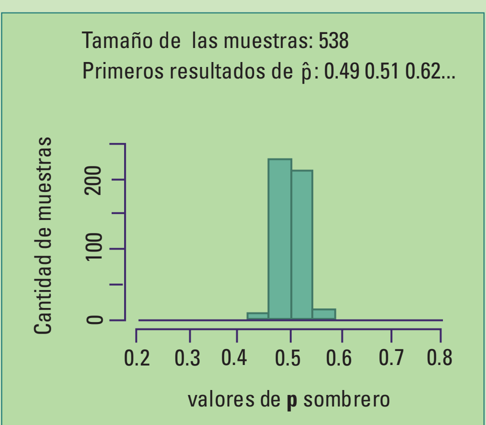


Figura 10.2.

En las figuras anteriores, en el eje horizontal se grafican los valores de las proporciones muestrales $\hat{p}$, las alturas de las barras muestran cuantas de las 1.000 muestras dieron valores dentro de cada uno de los grupos. Este tipo de gráfico se llama histograma (ver capítulo 16).

Vimos, en el capítulo 9, que la investigadora tomó una **única muestra** de 538 socios, y no sólo 35, obteniendo $\hat{p}$=0,51 ¿Qué pasa cuando tomamos muchas muestras de tamaño 538?

La figura 10.2 presenta los resultados de seleccionar **1.000 muestras** aleatorias simples **diferentes,** cada una de **tamaño 538** de una población para la cual la verdadera proporción es p=0,5. Las figuras 10.1 y 10.2 tienen los valores del eje horizontal en la misma escala, esto nos permite ver qué ocurre cuando el tamaño de las muestras se aumenta de 35 a 538: los valores están **más concentrados alrededor del valor verdadero 0,5** y, por lo tanto, **podemos confiar** más en un resultado que proviene de **una muestra de tamaño 538** que de una de tamaño 35.

En ambos casos, los valores de las proporciones muestrales $\hat{p}$ varían de muestra a muestra y están centrados en 0,5. Recordemos que **p=0,5** es el **valor verdadero** del **parámetro.** Algunas muestras tienen un $\hat{p}$ menor que 0,5 y otras mayor, sin que alguno de los dos sentidos esté favorecido. **El estimador de p ($\hat{p}$) no tiene sesgo;** esto ocurre tanto para muestras pequeñas como más grandes.

□ 10.2 Margen de error

Ya vimos que **los valores de $\hat{p}$** que provienen de las muestras de tamaño 35 **están más dispersos** que los de las muestras de tamaño 538 (figuras 10.1 y 10.2). Además, el 95% de las muestras de tamaño 538 dan estimaciones de p entre 0,4592 y 0,5408 o sea 0,0408 a cada lado del valor verdadero 0,5. Llamamos a 0,0408 **margen de error.**

Una **estimación de** p que resulte **de una muestra de tamaño 538** tendrá un **error de a lo sumo 0,0408** en el **95% de las muestras;** sólo el 5% tendrá un error mayor. Decimos que 0,0408 es **el margen de error de la estimación de p** con **un nivel de confianza del 95%.**

Como p=0,5 resulta un margen de error porcentual de 8,16%.

$$\text{Proporción de error:} \quad \frac{0{,}0408}{0{,}5} = 0{,}0816$$

$$\text{Error porcentual:} \quad \frac{100 \times 0{,}0408}{0{,}5} = 8{,}16$$

El margen de error del 8,16% significa que en el 95% de las veces que estimemos el parámetro p con un **tamaño de muestra 538 el error porcentual será menor a 8,16%.**

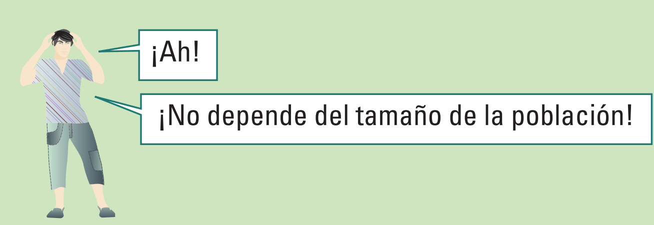
Para las **muestras de tamaño 35 el 95%** de los valores se encuentra entre 0,3429 y 0,6571 dando valores alejados del verdadero hasta una distancia de 0,1571 a cada lado. El **margen de error** porcentual, en este caso sería del **31,42 %.**

Hemos obtenido los márgenes de error tomando muchas muestras de una población a la que le conocíamos el valor verdadero del parámetro p. Este procedimiento es muy incómodo en situaciones reales.

Por suerte, los estadísticos han estudiado el problema de hallar el margen de error en general. Encontraron que cuando se utiliza una proporción muestral $\hat{p}$ calculada a partir de una muestra aleatoria simple de tamaño n para estimar una proporción poblacional p desconocida, con una confianza del 95%, **el margen de error** en un **muestreo aleatorio simple** será aproximadamente de $\frac{1}{\sqrt{n}}$. Al aumentar el tamaño de la muestra se reduce el margen de error.

Cuanto mayor sea el tamaño de la muestra, **MEJOR.**

El margen de error no depende del tamaño de la población, únicamente depende del tamaño de la muestra. Esto es cierto cuando la muestra es sólo una pequeña parte de la población, tal como ocurre en la mayoría de las encuestas. Una muestra aleatoria de tamaño 500 de una población de tamaño 100.000 es tan representativa como una muestra aleatoria de tamaño 500 de una población de tamaño 1.000.000.



Supongamos que se conversa con 20 alumnos/as de una escuela sobre la posibilidad de reducir las vacaciones de verano, agregando dos períodos de vacaciones uno en otoño y otro en primavera. Si a la mitad le pareció una buena idea, ¿estimaría que exactamente la misma proporción de todos los alumnos/as de la escuela está de acuerdo, suponiendo que la muestra es representativa de la opinión de todos?

Si el 50% de la muestra responde sí con una muestra de tamaño:	**El porcentaje de la población respondiendo sí podrá ser:**	
	Tan bajo como	Tan alto como
10	24%	76%
15	28%	72%
20	**31%**	**69%**
30	34%	66%
50	37%	63%
100	41%	59%
250	44%	56%
1.000	47%	36%

Con una muestra de tamaño 10, si 5 contestaron sí, podría ocurrir que el porcentaje verdadero de alumnos que quieren reducir las vacaciones de verano para agregar dos

vacaciones cortas en otoño y primavera sea tan baja como el 25% o tan alta como el 76%. Esta afirmación es correcta 95 de cada 100 veces. En toda la escuela el porcentaje de alumnos que quieren reducir las vacaciones de verano para agregar dos vacaciones cortas en otoño y primavera se encuentra entre el 24% y el 76%. Es un rango de valores muy amplio para el posible apoyo o no apoyo de la propuesta. Sería conveniente ampliar la muestra para tener un resultado más preciso.

¿Qué significa **"una confianza del 95%"**?

Significa que ese margen de error será válido el 95% de las veces que se calcule el estimador, **confiamos** que nos toque uno de los resultados buenos porque están en una relación de 95 a 5 con los resultados malos.

¿Qué significa **"margen de error"**?

El **margen de error** mide la diferencia máxima que se espera tener entre un resultado obtenido a partir de una muestra y su valor poblacional verdadero, el 95% de las veces.

□ 10.3 Error debido al muestreo aleatorio

Por más que una encuesta esté bien diseñada y bien conducida, dará el valor de un **estadístico** como estimación del **parámetro** poblacional. **Muestras diferentes darán valores diferentes** y el error debido al muestreo estará siempre presente. Pero podremos decir, con cierto grado de confianza, cuál va a ser la magnitud de ese error (denominado margen de error en la sección anterior). Se trata de **errores aleatorios,** surgen de utilizar una muestra en vez de la población total.

□ 10.4 Errores que no son debidos al muestreo aleatorio

Podemos llamar **equivocaciones** a algunos de estos errores. Pueden ocurrir en cualquier encuesta e incluso en los censos. Estas equivocaciones son posibles en todos los pasos, desde el registro del dato hasta obtención final del valor del estadístico. Actualmente, con el uso de procedimientos computarizados para muchos de los cálculos, se han reducido los errores de cálculo.

Otro tipo de errores son los debido a la presencia de sesgos en el muestreo, en las respuestas y/o en su registro (sección 6.3). Por ejemplo, pueden ocurrir cuando un **encuestado miente.** Lo llamamos sesgo de respuesta. Un respondente puede mentir respecto de su edad, de cuántas horas trabaja por día (puede pensar que trabaja poco y entonces las aumenta), de su salario (puede no querer que se sepa que gana mucho, o que gana poco) o puede haber olvidado cuantos paquetes de cigarrillos fumó la semana anterior.

Lo que no mide el margen de error: **No mide** el error que se comete debido al **sesgo** en el muestreo, ni el generado por las respuestas incorrectas y su registro. Estos pueden ser muy grandes, en comparación con el llamado margen de error y **no se reducen** al aumentar el tamaño de la muestra.

Podemos utilizar la analogía del juego del tiro al blanco para describir el efecto del sesgo y el tamaño de muestra en el error de muestreo. Supongamos que el centro del blanco (punto rojo de la figura 10.3) es el parámetro poblacional al que queremos acertar. Si estamos realizando un muestreo aleatorio, en cada muestra -es decir para cada tiro- obtendremos un punto cercano al centro. Algunas veces, el dardo caerá un poco arriba otras un poco abajo. Si en cambio el procedimiento tiene sesgo, los valores estarán todos desviados en una misma dirección. El esquema de la figura 10.3 muestra en la parte inferior puntos negros más concentrados que los de la parte superior, están representando un aumento en el tamaño de las muestras y una reducción de la variabilidad de los resultados. Sin embargo, el error la distancia de los puntos negros al rojo no se reduce al reducirse la variabilidad cuando hay sesgo (parte derecha del esquema).

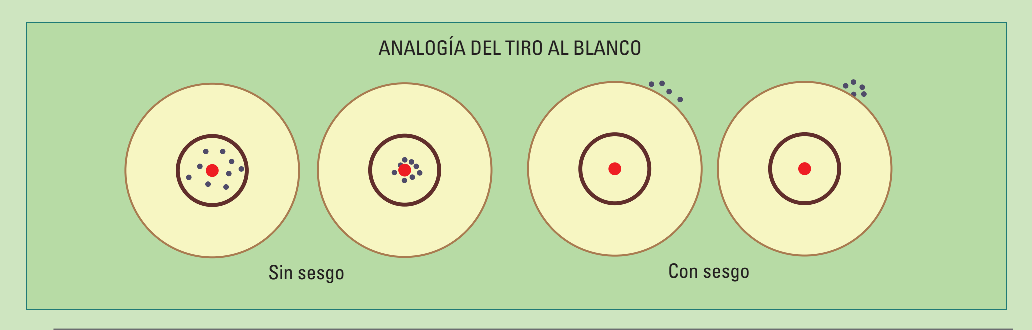


Figura 10.3. *Los puntos del panel inferior están más concentrados, los de la izquierda (representando un muestreo sin sesgo) están más cerca del punto rojo que los de la derecha.*

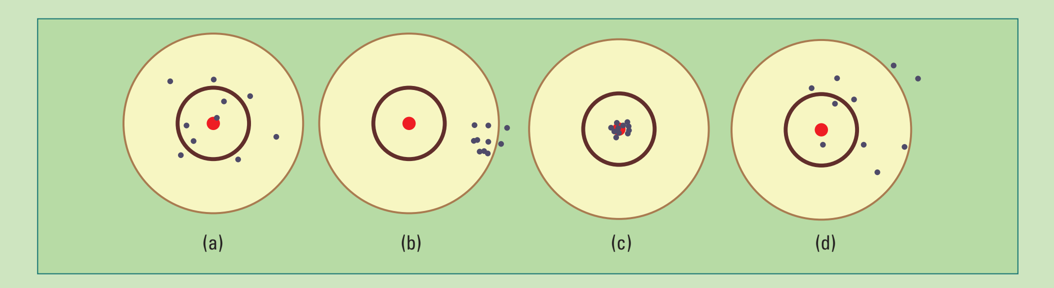
10.5 Actividades y ejercicios

1. Suponiendo que la verdadera proporción de socios a favor del candidato 1 del Club Grande de Fútbol fuera p=0,5.

   a) Si el tamaño de la muestra fuera 538,
   b) Si el tamaño de la muestra fuera 35,

   ¿Le sorprendería obtener 51% de socios a favor del candidato 1, y 37%? Para responder utilice los histogramas de 1.000 valores de $\hat{p}$ (figuras 10.1 y 10.2).

2. Siguiendo con la analogía del tiro al blanco de la sección 10.4, indique en cuál de las figuras siguientes los tiros son: I) precisos y sin sesgo, II) precisos y con sesgo, III) imprecisos y sin sesgo, IV) imprecisos y con sesgo.



3. La siguiente figura contiene gráficos semejantes a los de las figuras 10.1 y 10.2 para muchas repeticiones de distintos tipos de muestreos. Las alturas de las barras representan la frecuencia con la que apareció el valor del estadístico. El valor verdadero del parámetro está indicado. Agregue en cada gráfico a qué tipo de muestreo corresponde: I) preciso y sin sesgo, II) preciso y con sesgo, III) impreciso y sin sesgo, IV) impreciso y con sesgo.

Parámetro poblacional
(a)

Parámetro poblacional
(b)

Parámetro poblacional
(c)

Parámetro poblacional
(d)

4. Se eligen 3 alumnos para representar a su división en el centro de estudiantes. Si su división estuviera compuesta por 15 mujeres y 20 varones, y los representantes seleccionados fueran todos varones ¿Habría que sospechar de discriminación en contra de las mujeres?

   Veamos el comportamiento de la variabilidad muestral al elegir muestras pequeñas de una población pequeña (su división). Escriba los nombres de cada uno de los alumnos/as en papelitos del mismo tamaño y la misma forma. Coloque todos en una bolsa. Luego de mezclarlos, retire de la misma tres papelitos con los nombres de los alumnos seleccionados. Registre la cantidad de mujeres seleccionadas y devuelva los papelitos a la bolsa. Repita 25 veces. Construya un histograma como el de la figura 10.1 para la cantidad de mujeres seleccionadas en las 25 repeticiones. ¿Cuál es la cantidad promedio de mujeres en las 25 muestras?

5. Una encuesta nacional realizada a 437 varones y 1.125 mujeres obtuvo que al 64% de los varones les gustaría ver fútbol femenino por televisión, mientras que ese porcentaje se redujo a 42% entre las mujeres.

   a) Los encuestadores publicaron que el margen de error para una confianza del 95% es aproximadamente del 5% para los varones y 3 % para las mujeres. Explique a qué se debe esta diferencia.

   b) ¿Por qué es necesario incluir el margen de error al dar el resultado de una encuesta?

11. Estudios experimentales

El secreto está en la comparación. Grupo tratamiento versus grupo control.

Primero, una anécdota.

□ 11.1 La Dama del té

Al preparar un té con leche fría, ¿el sabor es el mismo al verter el té sobre la leche o la leche sobre el té?

Hacia fines de los años veinte (1920) en la ciudad de Cambridge (Inglaterra), una tarde de verano en una reunión de distinguidos académicos y sus esposas, una dama afirmaba: “el sabor no es el mismo”. Allí se encontraba Ronald Fisher, quien se entusiasmó en la discusión sobre si era posible saber si la dama podía, realmente, distinguir las dos formas de la preparación del té, únicamente por su sabor. Propuso presentarle varias tazas de té; algunas preparadas con el té vertido sobre la leche y otras con la leche vertida sobre el té. Varios asistentes a la reunión se unieron a la propuesta para ponerla en práctica vertiendo el té y la leche en distintos ordenamientos para que no pudiera adivinar. Así, una a una, fueron ofreciéndole las tazas de té, y registrando la respuesta de la dama sin realizar comentario alguno.



Ronald Aylmer Fisher (1890-1962) Matemático, estadístico, biólogo evolutivo y genetista inglés que estableció los cimientos de la estadística moderna. Statistical Methods for Research Workers (1925), The Genetical Theory of Natural Selection (1930), The design of experiments (1935), Statistical tables (1947).

Al diseñar el experimento se tratan de evitar los aciertos por casualidad. Si se le presenta una única taza y, simplemente adivina, su chance de acertar es 1 en 2. ¿Cuántas tazas son necesarias para reducir los aciertos casuales?

Ronald Fisher incluyó la anécdota en su libro “El diseño de los experimentos en 1935”. Mostró experimentos con diferentes diseños, para determinar si la Dama del Té podía detectar la diferencia, así como los cálculos de probabilidades de los aciertos por casualidad.

¿Pero qué pasó con la señora esa tarde? Dicen que acertó el orden de la preparación en todas las tazas.

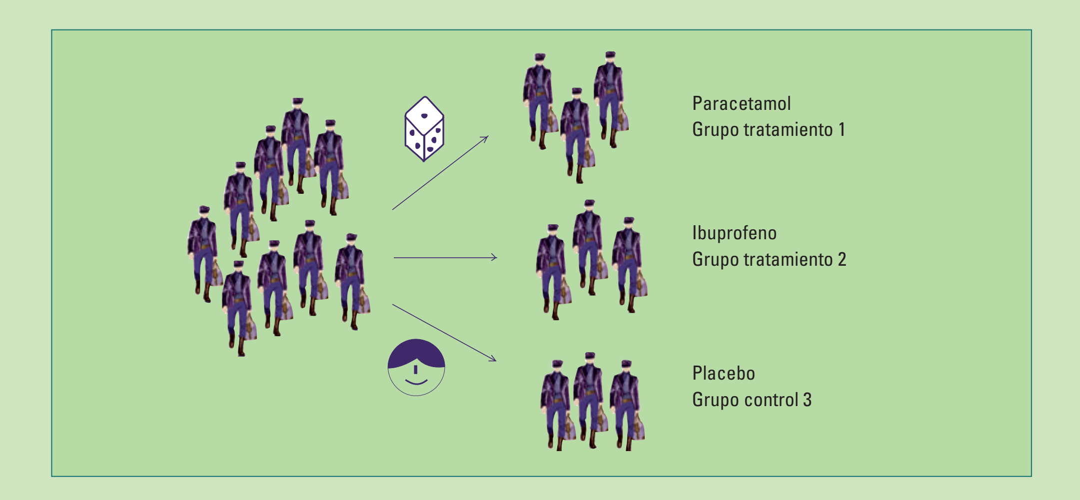
□ 11.2 Vocabulario

Cuando en un estudio se impone un cierto grado de control sobre los participantes y su entorno, como cuando se restringe una dieta o se administra una determinada dosis de un medicamento, se trata de un **estudio controlado.** Los participantes del estudio se asignan a dos o más grupos mediante un mecanismo aleatorio. Cada grupo recibe, por ejemplo, una dosis prefijada de una droga o diferentes drogas alternativas. Construimos así un **experimento comparativo aleatorizado,** llamado también simplemente **estudio controlado.**

El objetivo de la mayoría de los estudios experimentales es señalar una dirección causa-efecto entre dos variables. Se intenta resolver interrogantes del tipo: ¿cuál es la relación entre consumo de alcohol y problemas de visión? También como los siguientes:

- ¿Beber una copa de vino reduce la capacidad para conducir un automóvil?
- ¿Realizar actividad física mejora la fuerza de las mujeres de más de 55 años?
- ¿Tomar suplementos alimenticios con zinc ayuda a reducir la duración de un resfrío?
- ¿La letra con sangre entra?
- ¿La forma y la posición de su almohada modifican la calidad de su sueño?
- ¿La altura del taco de los zapatos influye en la comodidad del pie?

En estudios experimentales es frecuente comparar la efectividad de distintos tratamientos.



Consideremos un estudio comparando dos drogas con un placebo:

- **Participantes del estudio:** pacientes con osteoartritis.
- **Drogas:** acetaminofeno (también llamado Paracetamol o Tylenol), ibuprofeno.
- **Variable Respuesta:** grado de reducción del dolor de la rodilla y la cadera.

Los participantes se asignan a uno de tres grupos aleatoriamente. Dos grupos se llaman de tratamiento y uno grupo llamado de control. Los dos **grupos tratamiento** reciben Paracetamol e Ibuprofeno respectivamente y el **grupo control** recibe un placebo. En este estudio se pueden realizar preguntas como: ¿en cuál de los grupos la reducción del dolor ha sido la mayor? o ¿en cuál de los grupos se registra una menor proporción de eventos indeseables como, por ejemplo, gastrointestinales?

El experimento es aleatorizado cuando los sujetos han sido asignados a los distintos grupos mediante un mecanismo aleatorio.

El experimento es comparativo cuando sus conclusiones se obtienen comparando los resultados de los distintos grupos. Se espera que los grupos sólo difieran en la característica estudiada.

11.2.1 Grupo tratamiento versus grupo control

El grupo, o los grupos, tratamiento se componen de pacientes que reciben algún tratamiento. En nuestro ejemplo, teníamos dos grupos tratamiento uno asignado a Paracetamol y el otro a Ibuprofeno. En general, el grupo control se compone de individuos no tratados, o que reciben un tratamiento estándar bien conocido, cuyos resultados se compararán con el tratamiento nuevo.

11.2.2 Placebo

Un placebo es un tratamiento falso e inocuo, como una píldora de azúcar. A menudo se da a los miembros del grupo de control, para ocultar si están tomando el tratamiento (por ejemplo, Paracetamol) o no están recibiendo ningún tratamiento en absoluto.

El placebo es dado a los participantes asignados al grupo de control, precisamente con el fin de controlar el fenómeno llamado **efecto placebo:** los pacientes informan algún tipo de efecto cuando tienen la percepción de estar realizando un tratamiento, como tomar una píldora (aunque sea una píldora de azúcar).

El efecto informado puede ser positivo “Sí, me siento mejor”, o negativo “Me estoy sintiendo mareado”. Sin un placebo como referencia para hacer comparaciones, los investigadores no podrían distinguir si los resultados son debido al efecto real del tratamiento o al efecto placebo.

11.2.3 Ciego, doble ciego y triple ciego

En un **experimento ciego los participantes del estudio no saben si están en un grupo de tratamiento o en un grupo de control.** Un experimento a ciegas intenta eliminar cualquier sesgo de respuesta del sujeto debido a la información que recibe. Un paciente puede sentir una mejoría simplemente por saber que está tomando un medicamento de última generación muy bueno, o sentirse mal por saber que el medicamento es nuevo y no ha sido probado aún.

Un **experimento doble ciego** controla los posibles prejuicios tanto de los pacientes como de las/os investigadoras/es. **Ni los pacientes, ni las/os investigadoras/es conocen qué sujetos recibieron el tratamiento y cuáles no.** Un doble-ciego es mejor; los investigadores pueden tener un especial interés en los resultados (por algo están haciendo el estudio).

En un **experimento triple** ciego ni los pacientes, ni las/os investigadoras/es, ni las/os profesionales estadísticas/os que realizan el análisis de los datos pueden identificar a los sujetos con y sin tratamiento. Esto es lo mejor.

12. Estudios observacionales

Es necesario observar el mundo para intentar entender su funcionamiento.

La observación es el primer paso, a partir de ella se abrirán diversos caminos para desarrollar nuevas teorías y modelos; podrá motivar la realización posterior de nuevos estudios. Desde un punto de vista estadístico, los mejores resultados provendrán de estudios experimentales en comparación con estudios observacionales, pero algunas veces sólo es posible realizar estos últimos.

□ 12.1 Observar es bueno



Observando la naturaleza, Charles Darwin descubrió la selección natural como mecanismo de evolución de las especies.

En 1831 a los 22 años, emprendió un viaje alrededor del mundo de 5 años de duración como naturalista sin sueldo en un barco británico de reconocimiento, el velero Beagle. El primer indicio real respecto de la evolución de las especies no fueron los pinzones de las Islas Galápagos (1935), como se afirma muchas veces. Fue **tres años antes en la costa Argentina.** Ancló cerca de Bahía Blanca durante el primer año del viaje; desde allí llegó a Punta Alta y Monte Hermoso donde desenterró restos de fósiles de diversas criaturas, entre ellas encontró especies extintas ligeramente diferentes a las vivas. Le llamó la atención la presencia de un ñandú grande en las pampas y uno más pequeño el sur del río Negro.

A partir de las observaciones realizadas durante ese periplo, ya hacia 1838 Darwin tenía claro cómo la selección natural era un mecanismo de la evolución, aunque demoró la publicación de sus obras.

Consciente de las posibles repercusiones, y del rechazo de esa nueva visión de la realidad biológica por la conservadora sociedad victoriana, postergó su publicación, y decidió continuar añadiendo ocasionalmente nuevos datos.

"El Origen de las especies por selección natural" se puso a la venta recién a fines de 1859, agotándose ese mismo día. En enero de 1860 salió la segunda edición, llegó a tener seis ediciones en total durante la vida de Darwin.

La evolución es, 150 años después de su descubrimiento, tan firme como la "teoría" heliocéntrica (la Tierra gira alrededor del Sol) que también se desarrolló observando sin prejuicios (Copérnico, 1543). Cada una de estas teorías da una explicación confirmada, hasta cierto punto, por medio de la observación y la experimentación. A eso se refieren los científicos cuando hablan de una teoría.

Como ocurre con frecuencia con los avances de la ciencia, Darwin no fue el único en darse cuenta. También lo hizo Alfred Wallace, en forma independiente y simultánea; sus trabajos fueron presentados conjuntamente el 1 de julio de 1858, en la Linnean Society de Londres.

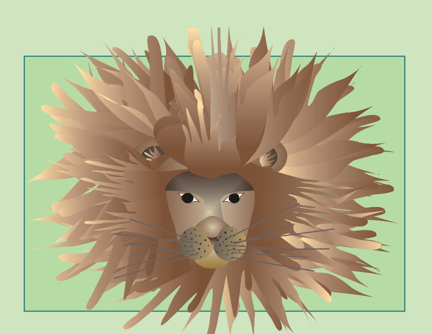
Las prácticas observacionales abren el camino para realizar experimentos. Con la evolución no es fácil experimentar porque en general se manifiesta luego de muchas generaciones.

Una manera de superar esta dificultad consiste en realizar experimentos utilizando especies con ciclos cortos de vida, por ejemplo, bacterias (500 generaciones en 75 días). Se las cultiva alterando alguna condición ambiental para observar la respuesta evolutiva. También se realizan experimentos con organismos superiores como moscas del género Drosophila; completan su ciclo en sólo 12 días, permitiendo detectar cambios generacionales en lapsos cortos y así estudiar su evolución.

Todos estos estudios requieren de la aplicación de técnicas estadísticas para obtener conclusiones con valor científico.

□ 12.2 Cuando sólo se puede observar

Imaginemos una investigación para conocer el comportamiento de los leones, en particular cómo las leonas enseñan a sus cachorros a cazar. Comienza por la observación. Al principio puede ser difícil saber qué registrar. Eventualmente, pueden surgir algunos patrones orientando las mediciones.



- ¿Con qué frecuencia cazan los leones?
- ¿Lo hacen los machos solos?
- ¿Lo hacen las hembras?
- ¿Van en grupos?
- ¿Los acompañan las crías, a partir de qué edad?
- ¿En qué etapa de la caza incorporan a los cachorros?

Las observaciones bien diseñadas, dirigidas a variables definidas claramente permitirán obtener resultados más convincentes.

En muchas oportunidades, no es ético realizar un estudio experimental. Por ejemplo, no es posible forzar a 100 personas a fumar 3 paquetes de cigarrillos por día y a otras 100 uno. En humanos sólo pueden realizarse estudios observacionales para responder preguntas como: ¿fumar provoca cáncer de pulmón?
Para evitar estas dificultades se conducen experimentos con animales, pero cada vez hay más reacciones contra este enfoque.

> En un **estudio observacional** se registran algunas características de individuos tratando de no influir en dichas mediciones.

Por ejemplo, se pueden considerar dos grupos de individuos, uno de sedentarios y otros de deportistas, -y sin influir en sus hábitos- se mide su nivel de colesterol en sangre, para evaluar si la actividad física lo afecta.

13. Estudio observacional versus estudio experimental

Algunas veces es posible realizar cualquiera de los dos tipos de estudio. En ese caso ¿cuál elegiríamos?

A continuación describiremos un ejemplo con ambas alternativas. Interesa estudiar si un suplemento diario de calcio en la dieta beneficia a las mujeres aumentando su masa ósea.

Diseño 1:
Se forma un primer grupo seleccionando consumidoras habituales de suplementos de calcio, y un segundo grupo con mujeres, también consumidoras de suplementos, pero sin calcio. Se mide la masa ósea en ambos grupos y se comparan los resultados. Se trata de un **diseño observacional** porque las mujeres **eligen libremente** tomar o no tomar suplementos de calcio.

Diseño 2:
Se selecciona un grupo de mujeres para participar del estudio. A la mitad de las mujeres, **se les asigna en forma aleatoria,** suplementos de calcio, a la otra mitad placebos con el mismo aspecto. Ni el médico ni la participante saben si ella pertenece al grupo de tratamiento o al grupo de control. Después de un tiempo de seguimiento del estudio, se comparan los dos grupos respecto a su masa ósea. Se trata de un **diseño experimental** porque las participantes son asignadas al azar en los grupos.

El enfoque experimental es más adecuado porque, por ejemplo, las mujeres que toman suplementos de calcio voluntariamente podrían ser precisamente las que mejor se cuidan en general y, por lo tanto tener mayor masa ósea por otras razones **(variables de confusión).** Con el diseño experimental, administrando en forma aleatoria el calcio a la mitad de las mujeres, se espera obtener grupos balaceados respecto de las variables que pueden afectar los resultados.

Como vimos en el capítulo 12, algunas veces no es posible realizar estudios experimentales. En estudios observacionales se puede controlar el efecto de los factores de confusión realizando las comparaciones en subgrupos más pequeños y más homogéneos. Por ejemplo, en un estudio sobre el efecto del tabaquismo en la salud, se comparan fumadores con no fumadores dentro de subgrupos con edad similar, el mismo género y características similares en todos los posibles factores influyentes en la patología en estudio además del tabaquismo. Se trata de lograr homogeneidad dentro de los subgrupos excepto por la condición en estudio, en este caso el tabaquismo.

13.1 Actividades y ejercicios

1. ¿Cuáles de las siguientes afirmaciones son verdaderas?

   a) En un estudio experimental un grupo es forzado a seguir un tratamiento con el propósito de observar una respuesta.
   b) En un estudio observacional se recoge información sin realizar ninguna acción para modificar la situación existente.
   c) Las encuestas son estudios observacionales, no son experimentos.

2. ¿Cuáles de las siguientes afirmaciones son verdaderas?

   a) En un experimento los investigadores deciden como se colocan a las personas en los distintos grupos.
   b) En los estudios observacionales, los participantes eligen en qué grupo estar.
   c) Un grupo control generalmente es de elección voluntaria.

3. En un estudio para determinar el efecto de la actividad física en el nivel de colesterol, se comparó el nivel de colesterol de 100 sujetos que concurrían al gimnasio 4 veces por semana con 100 sujetos que no realizaban actividad física. En un segundo estudio, 50 sujetos fueron asignados aleatoriamente para asistir al gimnasio 4 veces por semana y otros 50 para participar en clases de pintura. Indique cuales de las siguientes afirmaciones son verdaderas.

   a) El primer estudio es un experimento controlado, el segundo es un estudio observacional.
   b) El primer estudio es un estudio observacional, el segundo es un experimento controlado.
   c) Ambos son experimentos controlados.
   d) Ambos son estudios observacionales.
   e) Cada estudio es un poco experimental y un poco observacional.

4. En un estudio para determinar el efecto de la provisión gratuita de leche a los niños de las escuelas de un distrito, se asignó a las escuelas en forma aleatoria uno o dos litros de leche por semana por alumno, y se registraron los días de ausencias por enfermedad durante un año. En otro estudio realizado en un hospital de niños se preguntó, mediante un cuestionario entregado en la sala de espera, cuánta leche tomaba el niño por semana y cuántos días había faltado a la escuela por enfermedad en el último año. Indique cuáles de las siguientes afirmaciones son verdaderas.

a) El primer estudio es un estudio experimental sin grupo control y el segundo es un estudio observacional.
b) El primer estudio es un estudio observacional y el segundo es un estudio experimental controlado.
c) Ambos estudios son observacionales.
d) Ambos estudios son experimentos controlados.

5. Se seleccionaron al azar 10 de 20 personas que sufrían dolor de cabeza y se les dio chocolates con sabor a menta y color modificado para ocultar el chocolate. A los otros 10 se les dio tabletas de aspecto y gusto similar pero sin chocolate. Al día siguiente 6 de las 10 personas que habían consumido chocolate reportaron haber sufrido dolor de cabeza; ninguna de las que no recibieron chocolate reportaron dolor de cabeza.

a) Se trata de un estudio observacional para determinar el efecto del chocolate sobre el dolor de cabeza.
b) Se trata de una encuesta en la cual se eligieron 10 de 20 personas con dolor de cabeza para darles tabletas de chocolate con sabor a menta.
c) Se trata de un censo de 20 personas que suelen tener dolor de cabeza; se registró a cuántas personas se les dio chocolate y cuántas tuvieron dolor de cabeza.
d) Se realizó un estudio utilizando el chocolate como placebo para estudiar una causa del dolor de cabeza.
e) Se realizó un experimento en el cual al grupo tratamiento se le dio chocolate y al grupo control no.

6. Indique cuales de las siguientes afirmaciones son verdaderas.Interesa saber cuántos varones y cuántas mujeres asisten a una escuela determinada. ¿Cuál es la forma más adecuada de obtener esa información? Mediante un

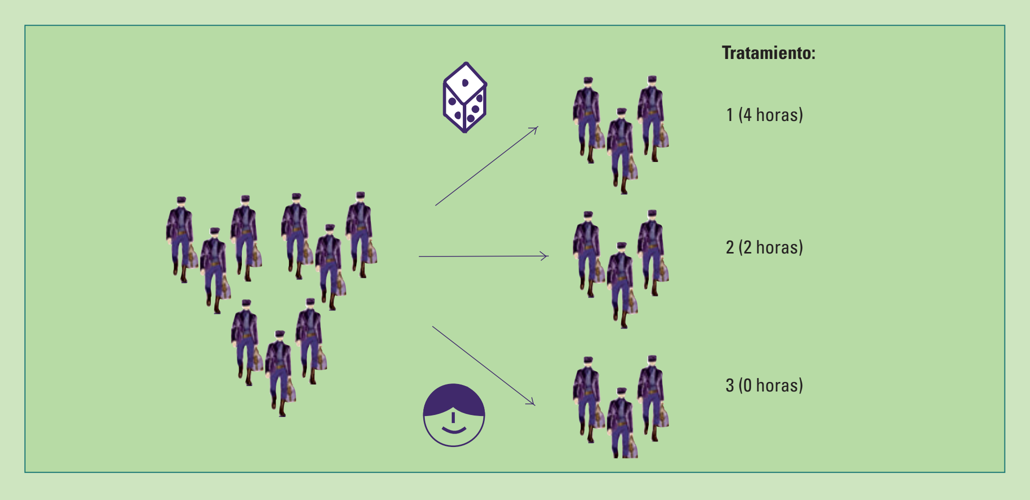
a) Censo
b) Encuesta
c) Experimento controlado
d) Estudio observacional

14. No siempre los tratamientos son tratamientos

En jerga estadística los tratamientos no se usan no sólo en medicina.

Cuando en un estudio interesa el efecto de un único factor **(variable explicativa)** sobre una **variable respuesta** se denomina **tratamiento** a cada uno de los **niveles del factor.**

Comencemos con un ejemplo: interesa evaluar si la asistencia a clases de apoyo los días sábados **(el factor)** puede influir en el rendimiento de los estudiantes **(la respuesta).** Para ello se puede **dividir a los alumnos** seleccionados en **tres grupos.** El primero con 4 horas por semana de clases de apoyo (tratamiento 1), el segundo dos horas (tratamiento 2) y el tercero ninguna (tratamiento 3). Al final del trimestre los alumnos darán una prueba con el fin de comparar las respuestas a los tratamientos.



Es importante que la asignación de los alumnos a los grupos se haga al azar y no en forma voluntaria. El experimento será ciego si los evaluadores ignoran a qué grupo pertenece cada alumno, y nunca será doble ciego por que los alumnos siempre sabrán cuantas horas toman de clases.

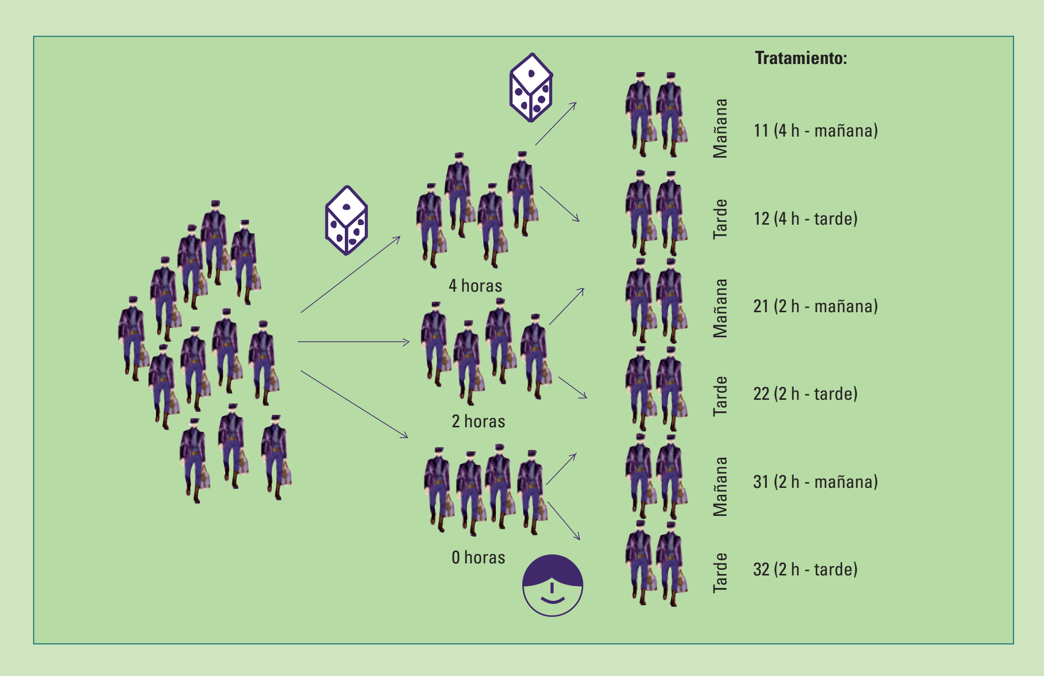
En el ejemplo, el factor es la asistencia a clases de apoyo medido en horas, y la variable respuesta el desempeño del alumno al final del trimestre. Tenemos tres tratamientos, uno para cada nivel del factor (4 horas, 2 horas, 0 horas).

Estamos utilizando el término tratamiento en sentido amplio. En estadística nos referimos a tratamientos no solo a los de medicina, sino también en referencia a distintas áreas, cuando se quiere comparar el efecto de cierto tipo de intervención sobre alguna respuesta. Más precisamente se trata de factores (variables explicativas) que podrían tener efecto sobre variables de respuesta. Un grupo se "trata" con algún nivel de la variable explicativa y el resultado a medir es de la variable respuesta.

Compliquemos un poco el ejemplo. Supongamos que los alumnos fueran a su vez divididos, aleatoriamente, en clases de apoyo en dos turnos: mañana y tarde.

Cuando hay más de un factor en estudio, se denomina tratamiento a cada una de las combinaciones de los diferentes niveles de cada uno de los factores. En este caso, como tenemos dos factores, uno con tres niveles (4, 2, y 0 horas) y otro con dos (mañana y tarde), resulta un total de 6 tratamientos:

- Tratamiento 11: 4 horas - mañana
- Tratamiento 12: 4 horas - tarde
- Tratamiento 21: 2 horas - mañana
- Tratamiento 22: 2 horas - tarde
- Tratamiento 31: 0 horas - mañana
- Tratamiento 32: 0 horas - tarde



En el ejemplo sobre la efectividad de las clases de apoyo se ilustra el uso del término "tratamiento". Es un estudio experimental; pero, este concepto puede aplicarse también a estudios observacionales. Cuando se comparan sujetos que concurren habitualmente al gimnasio 4 veces por semana con sujetos que no realizan actividad física, tenemos un estudio observacional; el factor en estudio es la actividad física y los "tratamientos" son 2:

- concurre 4 veces por semana al gimnasio
- no realizan actividad física.

15. Mediciones válidas

Una variable es una medida válida de un concepto si lo representa adecuadamente.

Medir una característica de un individuo (persona, objeto, animal, etc.) significa asignarle un número expresable en distintas unidades que la represente. El resultado de esa medición es **variable** y toma diferentes valores dependiendo del individuo a quien se le está realizando la medición.

Algunas características, como el peso y la talla, pueden ser más sencillas de medir que otras como la inteligencia o la percepción del dolor.

Muchas veces disponemos de un **instrumento** para realizar la medición. Para obtener la longitud de una mesa utilizamos una cinta métrica; estará expresada en centímetros para la UE y Argentina y en pulgadas para Estados Unidos, es decir, como unidades para expresar mediciones se pueden utilizar centímetros ó pulgadas.

La medición requiere de:

- Un proceso previo de **transformación de conceptos** (longitud, desempleo, dolor, nivel socioeconómico, etc.) en variables definidas con precisión.
- La elección del **instrumento** para medirlas.

La utilización de una cinta métrica para transformar la idea "longitud" en un número es directa, porque sabemos exactamente qué queremos decir con longitud, pero en otros casos puede resultar mucho más complicado. Para medir la inteligencia se requiere de un cuestionario y un mecanismo de cálculo para obtener un número de acuerdo con las respuestas.

Muchas veces no disponemos de mecanismos o instrumentos de medición adecuados para abordar un tema de nuestro interés, pero podemos utilizar resultados de organismos del estado o empresas privadas. Al utilizarlos es importante evaluar cómo están definidos, y si se trata de medidas válidas para describir las propiedades que se pretenden medir.

En la próxima sección presentamos la cantidad de accidentes de tránsito y la cantidad de desocupados como ejemplo de dos conceptos definidos precisamente para obtener, sin demasiadas dificultades, mediciones válidas. Veremos más adelante que a veces no es posible obtener mediciones claramente válidas (dolor, inteligencia). También, que para un mismo concepto se puede obtener más de una medición válida. Finalmente, ilustramos con más detalle la construcción de un instrumento para medir la evolución de los precios al consumidor.

◻ 15.1. Sin demasiadas dificultades

15.1.1. Accidentes de tránsito

¿Cómo se mide la seguridad en las rutas? Se puede contar la cantidad de víctimas fatales por año en el momento de un accidente de tránsito. En nuestro país el Registro Nacional de Antecedentes de Tránsito (ReNAT) publica esa información. Como vimos en la capítulo 3, se puede utilizar esa **cantidad de muertes** como **variable** para medir la seguridad en las rutas; también la tasa de víctimas fatales por cada millón de habitantes o por cada cien mil vehículos circulantes.

Podríamos utilizar los datos ReNAT sin averiguar de qué manera se elaboran. Sin embargo, como consumidores de la información, deberíamos indagar un poco más. Para contar muertes fatales en las rutas es necesario saber exactamente a qué se refiere el término "víctimas fatales". ¿Se trata de peatones atropellados por un auto?, ¿automovilistas arrollados por un tren? Contestar estas y otras preguntas, permite saber qué se está contando. ¿Se incluyen los fallecidos dentro de las 24 h del accidente, o dentro del primer mes, etc.?

15.1.2. Desocupación

La Encuesta Permanente de Hogares (EPH, INDEC) releva información para calcular trimestralmente la tasa de desocupación.

Para ser desocupado, es necesario formar parte del mercado laboral e integrar la población económicamente activa .La población económicamente activa es la que se cuenta en el denominador. Entre ella los que están buscando trabajo van en el numerador, queda claro en la definición. Esta población está formada por las personas con alguna actividad económica o que sin tenerla la están buscando activamente. Los estudiantes o los jubilados, no deben contarse como parte de los desocupados aunque no tengan empleo, porque no están disponibles para realizar un trabajo.

La EPH define la población desocupada como el conjunto de todas las personas que, no teniendo ocupación, están buscando activamente trabajo. Este concepto no incluye a los desocupados que han suspendido la búsqueda por falta de oportunidades visibles de empleo, ni a los subocupados involuntarios. Es decir, una persona para ser desocupada debe estar disponible para un trabajo y buscando uno. La tasa de desocupación daría diferente si se utilizara otra definición.

Tasa de desocupación: Es la relación entre la población desocupada y la población económicamente activa, expresada en porcentaje.

$$\text{tasa de desocupación} = 100 \times \frac{\text{cantidad de personas desocupadas}}{\text{cantidad de personas económicamente activas}}$$

Importan mucho los detalles. Los resultados también dependen de cómo se realizan las preguntas para obtener la información relacionada con la desocupación. No se trata simplemente de preguntarle al entrevistado: " ¿Forma parte del mercado laboral?" "¿Está desocupado?"

Se necesitan muchas preguntas para clasificar a una persona en empleada, subempleada o no perteneciente al mercado laboral. De eso se encarga la EPH. Ahora veamos algunos de sus resultados. En particular consideraremos **la tasa de desocupación desde 1995 hasta 2008** y la presentaremos en un **gráfico de tiempo.**

Gráfico de tiempo: Este tipo de gráfico se utiliza para examinar la evolución a lo largo del tiempo de alguna variable. Tiene una unidad de tiempo en el eje horizontal (como meses o años) y en el eje vertical alguna cantidad (ingresos de los hogares, tasa de natalidad, ventas totales, porcentaje de la gente en favor del presidente, y así sucesivamente). En cada período de tiempo, la cantidad está representada por un punto, y los puntos están conectados por líneas.

Los datos correspondientes a la tasa de desocupación deben dividirse en dos períodos. El primero de 1995 hasta 2002 - porque el índice se publicaba 2 veces al año (en mayo y octubre)- y el segundo de 2003 hasta el 2008, porque la publicación se realiza 4 veces al año. En este segundo período, con inicio en enero de 2003, la EPH introdujo mejoras metodológicas. El INDEC **cambió el instrumento de medición** del mercado laboral para mejorar la calidad de la información.

La figura 15.1 muestra la evolución de la tasa de desempleo con un quiebre entre la medición de octubre de 2002 y la del 1er trimestre del 2003, poniendo de manifiesto el cambio de la metodología, Las mediciones de los dos períodos no son comparables directamente.

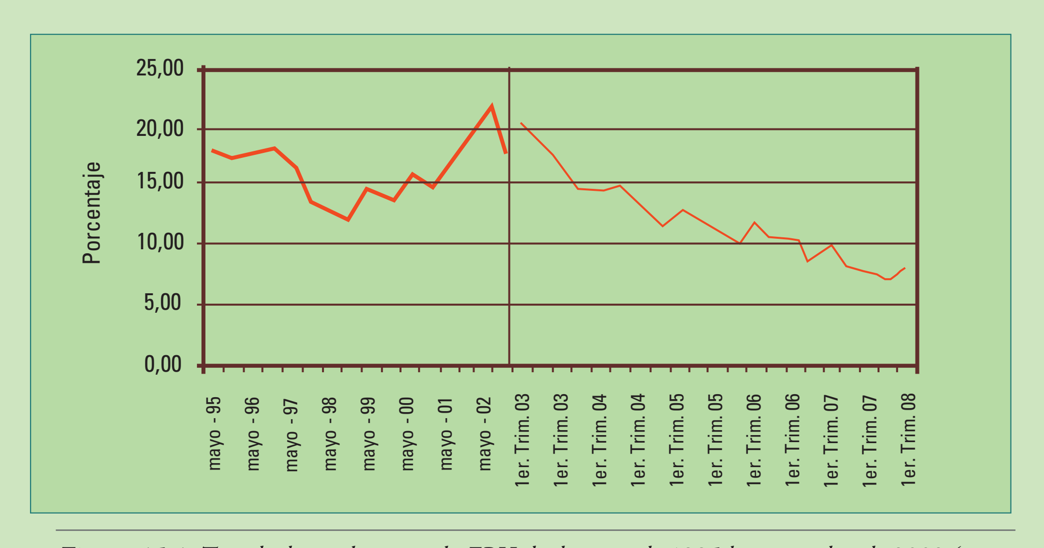


Figura 15.1. *Tasa de desempleo según la EPH desde mayo de 1995 hasta octubre de 2002 (con periodicidad bianual, mayo, octubre) y desde el 1er trimestre del 2003 hasta el 1er. trimestre de 2008 (con periodicidad trimestral).*

Vemos una tendencia decreciente de la desocupación desde mayo de 1995 hasta octubre de 1998, a partir de allí la tasa de desocupación empieza a aumentar llegando a un máximo en mayo de 2002. Luego se observa una tendencia decreciente en la tasa de desocupación hasta el 1er. trimestre de 2008.

Es importante destacar que la escala temporal en los dos períodos no es la misma. No son visualmente comparables las tendencias crecientes y decrecientes entre ellos.

Los cambios metodológicos incluyen mejoras de los formularios, ampliación de la muestra, aumento de la frecuencia con que se recoge la información, incorporación de procedimientos digitales, etc.; son habituales en todos los institutos de estadística del mundo. La comparación de los resultados obtenidos con metodologías diferentes es más difícil.

La figura 15.2 muestra un gráfico de tiempo para la tasa de desocupación de USA desde 1979 hasta 1996. Aquí vemos tres cortes, porque en 1986, 1990 y1994 se produjeron cambios metodológicos en el relevamiento y procesamiento de la información (http://www.bls.gov/cps/eetech_methods.pdf).

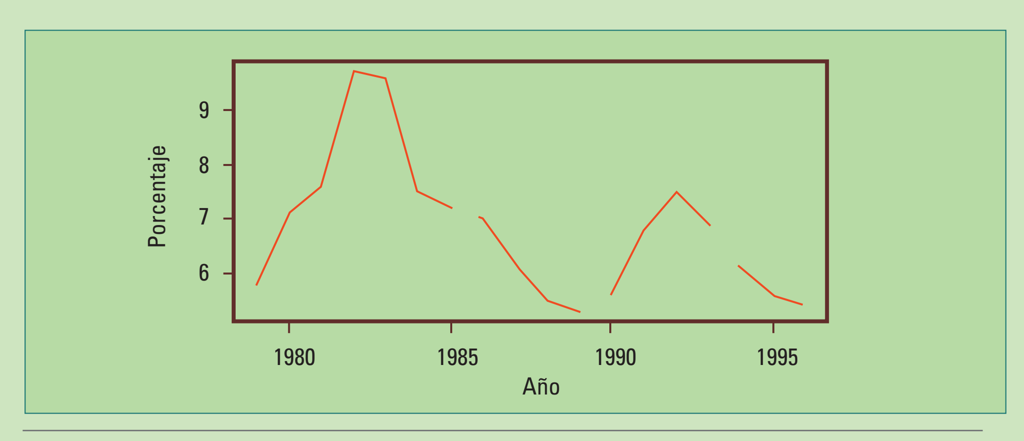


Figura 15.2. *Tasa de desempleo de USA ftp://ftp.bls.gov/pub/suppl/empsit.cpseea1.txt.*

15.2. Puede ser más difícil

Nadie pondría objeciones en el uso de una cinta métrica para medir la longitud de una mesa, sin embargo, algunas personas se pueden oponer a exigir una prueba de evaluación para decidir si un alumno está capacitado para ingresar a una universidad. Aunque todos coincidirán en que es una mala idea medirles la altura a los aspirantes y aceptar a los más altos. ¿Por qué? Porque la altura no tiene nada que ver con estar o no estar preparado para la Universidad.

> Una **variable** es una medida **válida** de un concepto si lo representa adecuadamente, o es una característica importante.

¿Está midiendo lo que le interesa medir? Si es así la variable es válida; si no, no lo es.

La validez, en términos generales, se refiere al grado en que una variable representa realmente la característica a medir. Por ejemplo, si interesa medir la inteligencia y se utiliza la memoria como medida, esta no es válida.

Es válido medir la altura con una cinta métrica, pero no es válido utilizar la altura como medida de la capacidad de un aspirante a ingresar a la universidad.

Muchas veces el problema de la validez de una variable para medir un concepto se encuentra en la naturaleza misma de ese concepto, tal como ocurre con la inteligencia.

¿Qué es la inteligencia? ¿El cociente intelectual mide la inteligencia? Algunos psicólogos dirán sí. Otros argumentarán que la inteligencia está compuesta por una gran variedad de capacidades mentales y por lo tanto no puede ser medida con un único instrumento. Si no podemos decidir qué es exactamente la inteligencia, menos podremos decidir cómo medirla.

Otro ejemplo problemático es la medición del dolor. El dolor es una experiencia personal. La forma más común de medirlo es preguntarle al paciente sobre las dificultades que ese dolor le acarrea. También se le puede pedir una descripción del nivel de dolor en alguna escala, ésta no significará lo mismo para diferentes personas.

Aún así, a veces no es tan difícil hallar la variable que provee una medición válida.

La **cantidad de accidentes fatales** por año no es una variable válida si queremos evaluar los resultados de una campaña de educación vial, pues los accidentes pueden aumentar si se incrementa el parque automotor o aumenta la población, como vimos en la sección 3.1.2, la **tasa de muertes** por cada 100.000 vehículos en circulación es una medida más adecuada.

Muchas veces una **tasa** (dada como fracción, proporción o porcentaje) es la medida válida, en contraposición con tomar simplemente **cantidades.**

15.3. Más de una válida

Una variable es una medida válida de un concepto si lo representa adecuadamente o es una característica importante de dicha propiedad. Pero puede haber más de una medida válida para un mismo concepto.

Un aviso institucional televisivo anuncia “1 de cada 8 mujeres, puede padecer cáncer de mama en algún momento de su vida, pero las mujeres que tienen antecedentes familiares tienen 2 a 4 veces más riesgo”. Pero ¿cómo se mide el riesgo? Veremos dos maneras de medir el riesgo y como se calculan.

La **proporción** y la **razón** son dos medidas válidas para medir el riesgo de padecer una enfermedad.

Veamos primero cómo se calculan esas dos medidas y luego, cómo los resultados pueden ser distintos:

- Cuando se mide como una **proporción** se toma la cantidad de personas que experimentan el suceso (padecer cáncer de mama) y se lo divide por la **cantidad total** de personas en riesgo de tener el evento.
- Cuando se mide como una razón también se toma la cantidad de personas que experimentan el suceso (padecer cáncer de mama) pero en este caso se divide solamente por la **cantidad de personas que no experimentan** el suceso.

De acuerdo al aviso, el riesgo de padecer cáncer de mama para la población es:

- Cuando se mide como una **proporción:** $\frac{1}{8} = 0{,}125$

- Cuando se mide como una **razón:** $\frac{1}{8} = 0{,}143$

Una proporción de 0,125 y una razón de 0,143 son dos medidas válidas del mismo riesgo.

Siguiendo con el aviso: ¿qué significa tener 4 veces más riesgo de padecer cáncer de mama? No es lo mismo cuadruplicar la proporción que cuadruplicar la razón:

Cuadruplicar la proporción resulta en una proporción de $\frac{4}{8} = \frac{1}{2}$
La mitad de las mujeres padecerán cáncer y la otra no.

Cuadruplicar la razón resulta en una razón de $\frac{4}{7}$
4 padecerá la enfermedad y 7 no.

El aviso no aclara qué medida se utilizó.

15.4. Números índices

Los **números índices** se utilizan, en forma similar a la tasa de desocupación (sección 15.1), para mostrar cómo cambia una característica con en el tiempo. Describen el cambio porcentual respecto al valor en un período base.

Tienen la ventaja de ser adimensionales, es decir, no tienen unidades. Por ejemplo, si se trata de un índice para reflejar la evolución de la superficie cubierta construida por mes, no importará si esa superficie está medida en metros cuadrados o en pies cuadrados.

Un número índice es el cociente, entre el valor de una variable en un momento del tiempo y el valor de la misma variable en otro momento llamado período base, multiplicado por 100:

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$



¡Un número índice es un porcentaje!

PRECIOS PROMEDIO, MENSUALES POR LITRO DE NAFTA SÚPER EN SURTIDOR. AGO-99 A JUL-00 TABLA 15.1

Período	Precio ($/litro)
Ago-99	0,920
Sep-99	0,940
Oct-99	0,975
Nov-99	0,977
Dic-99	1,024
Ene-00	1,049
Feb-00	1,051
May-00	1,055
Jun-00	1,052
Jul-00	1,064

Secretaría de Energía. Boletín de Precios de Combustibles – junio de 2001

Veamos un ejemplo sencillo. La Secretaría de Energía releva el precio por litro de nafta súper en surtidor en 500 estaciones de servicio distribuidas por todo el país e informa mensualmente el promedio de esos precios.

El precio promedio mensual en 500 estaciones de servicio, por litro de nafta súper en surtidor, se muestra en la segunda columna de la tabla 15.1. Un litro de nafta costaba $ 0,92 en agosto de 1999 y $ 1,064 en julio de 2000. Utilizaremos los datos de la tabla 15.1 para ilustrar cómo se calculan los números índices

Tomando como período base el mes de agosto de 1999, el número índice del precio de la nafta en julio de 2000 es 115,65:

Los **números índices** se utilizan, en forma similar a la tasa de desocupación (sección 15.1), para mostrar cómo cambia una característica con en el tiempo. Describen el cambio porcentual respecto al valor en un período base.

Tienen la ventaja de ser adimensionales, es decir, no tienen unidades. Por ejemplo, si se trata de un índice para reflejar la evolución de la superficie cubierta construida por mes, no importará si esa superficie está medida en metros cuadrados o en pies cuadrados.

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{1{,}064}{0{,}92} \times 100$$

$$= 115{,}65$$

El índice del precio de la nafta en el período base (agosto de 1999) es 100:

$$\text{número índice} = \frac{0{,}92}{0{,}92} \times 100$$
$$= 100$$

Por supuesto, los valores del índice dependen del período tomado como base.

Tomando como período base el mes de diciembre de 1999, el número índice del precio de la nafta en julio de 2000 es 103,91:

$$\text{número índice} = \frac{1{,}064}{1{,}024} \times 100$$
$$= 103{,}91$$

El **número índice** de una variable indica el valor de esa variable como **porcentaje** del valor en el **período base.**

PRECIOS PROMEDIO E ÍNDICES MENSUALES, POR LITRO DE NAFTA SÚPER EN SURTIDOR, AGOSTO DE 1999 A JULIO DE 2000 TABLA 15.2

Período	Precio ($/litro)	Índice Ago 99 = 100	Índice Dic 99 = 100
Ago-99	0,920	100,00	89,84
Sep-99	0,940	102,17	**91,80**
Oct-99	0,975	105,98	95,21
Nov-99	0,977	106,20	95,41
Dic-99	1,024	111,30	100,00
Ene-00	1,049	114,02	102,44
Feb-00	1,051	114,24	102,64
May-00	1,055	114,67	103,03
Jun-00	1,052	114,35	102,73
Jul-00	1,064	**115,65**	103,91

La tabla 15.2 muestra los precios promedio en surtidor de nafta súper junto con los números índices calculados con los datos de la tabla 15.1. En la columna 3 aparecen los índices con agosto de 1999 como período base y en la columna 4 el período base es diciembre de 1999.

El número índice 115,65 significa que el precio promedio de la nafta súper en julio de 2000 era 115,65% del valor base, es decir, el incremento respecto del valor base (agosto de 1999) es del 15,65%.

El número índice 91,80 de septiembre 1999 significa que en ese mes el valor era el 91,80% del valor base, o sea un 8,20 % menor que en el período base (diciembre de 1999).

El número índice para el período base es 100; por ejemplo si el período base es el mes agosto de 1999 se indica como "**agosto de 1999 = 100".**

Conocer el **período base** es esencial para poder interpretar un número índice.

Muchas veces se utiliza como período base un año. Por ejemplo, si el período base para un índice de precios de un litro de nafta es el año 1999 se lo indica como "1999 = 100" y el valor base es el promedio de los precios mensuales. En este caso, salvo cuando el precio promedio de algún mes coincida con el promedio anual, ningún mes tendrá índice 100.



El precio promedio de la nafta, entre las estaciones de servicio, no coincide con el gasto promedio en nafta de una persona.

15.4.1. Índice de precios al consumidor

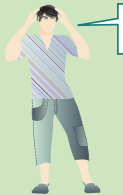
El índice de precios al consumidor es uno de los indicadores más importantes generados por los institutos de estadísticas oficiales del mundo. Es una medida del poder de compra de la unidad monetaria, pesos en nuestro caso. Afecta las decisiones gubernamentales y está vinculado directamente con gran parte de la economía.

Inquilinos y propietarios comparan su evolución con la pactada en los contratos de alquiler, para ver quién gana y quién pierde.

Pero ... ¿Qué es el índice de precios al consumidor?

El índice de precios al consumidor es un indicador de la evolución en el tiempo, en relación a **un período base,** de los precios de la canasta familiar. La **canasta familiar** es un grupo prefijado de bienes y de servicios representativos del gasto de los hogares en una **zona de referencia.** La evolución de los precios al consumidor puede ser diferente entre provincias.

Por ejemplo, un período base puede ser el año 1999 y la zona de referencia el Gran Buenos Aires.



¿Período base? ¿Canasta familiar? ¿Zona de referencia?

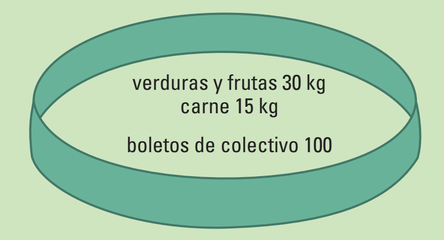
Se registran los precios de las componentes de la canasta familiar (bienes y servicios) en un período inicial o período base. Los precios de ese momento son "la base" del índice. Luego, se siguen registrando los precios a lo largo del tiempo de **la misma canasta familiar.** Para calcular el índice se comparan los precios de cada período con los precios del período base. Pero, ¿cómo se hace esa comparación?

Para contestar esa pregunta empecemos por un ejemplo. Consideremos los gastos de una familia hipotética (es decir, inventada) la familia Pérez cuya canasta familiar mensual tiene solamente 3 componentes: verduras y frutas (30 kg), Carne (15 kg), Boleto colectivo (100 viajes) (esto también es hipotético).

Estructura de gastos familia Pérez julio de 2008 (período base). Tabla 15.3

Bienes y servicios	Cantidad (julio-08)	Precio por unidad (julio-08)	Costo (julio-08)
Verduras y frutas	30 kg	$ 10/kg	$ 300 = $ (30 x 10)
Carne	15 kg	$ 20/kg	$ 300 = $ (15 x 20)
Boleto de colectivo	100 viajes	$ 1/viaje	$ 100 = $ (100 x 1)
Costo total			$ 700

La tabla muestra los costos de los bienes y servicios de la familia Pérez en el período base. En la primera columna se muestran las cantidades de cada uno de los bienes y servicios. En la segunda su precio por unidad, y en la tercera el costo total, que se obtiene multiplicando el precio unitario por la cantidad de unidades. Finalmente, el gasto total de la familia en el mes de Julio de 2008 resulta de sumar el gasto en cada uno de los bienes y servicios ($ 700).



¿Cuánto cuesta ésta canasta?

Para hallar el valor del índice en el mes de agosto de 2008, para la familia Pérez, utilizamos los precios por unidad correspondientes a ese mes (tabla 15.4, son valores inventados para ejemplificar el cálculo), con la misma cantidad de bienes y servicios de los del período base (julio de 2008).

Costos del mes de agosto de 2008. Tabla 15.4

Bienes y servicios	Cantidad (julio-08)	Precio por unidad (agosto-08)	Costo (Agosto-08)
Verduras y frutas	30 kg	$ 11/kg	$ 330 = $ (30 x 11)
Carne	15 kg	$ 20/kg	$ 300 = $ (15 x 20)
Boleto de colectivo	100 viajes	$ 1,30/viaje	$ 130 = $ (100 x 1,3)
Costo total			$ 760

Los mismos bienes y servicios costaban $ 700 en el mes de julio de 2008 y en el mes de agosto $ 760. Por lo tanto el número índice, para la familia Pérez, agosto de 2008 (julio de 2008 =100) es 108,57:

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{760}{700} \times 100$$

$$= 108{,}57$$

El índice mide la variación del valor de la canasta familiar respecto del período base. Para su cálculo se debe registrar el costo de la misma colección de bienes y servicios, **las mismas cantidades y los mismos productos** del período base. En el cálculo no interviene el cambio de hábitos posiblemente introducido por la familia al producirse, por ejemplo, un 30% de aumento en el costo de un viaje.

Describimos a continuación **otra forma de cálculo** del número índice. Muestra en forma explícita cómo los precios por unidad de cada producto (tablas 15.3 y 15.4) ingresan con **ponderaciones fijas:**

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{330 + 300 + 130}{300 + 300 + 100} \times 100$$

$$= \frac{11 \times 30 + 20 \times 15 + 1{,}3 \times 100}{700} \times 100$$

$$= \frac{\frac{11}{10} \times 300 + \frac{20}{20} \times 300 + \frac{1{,}3}{1} \times 100}{700} \times 100$$

$$= \frac{11}{10} \times \frac{300}{700} \times 100 + \frac{20}{20} \times \frac{300}{700} \times 100 + \frac{1{,}3}{1} \times \frac{100}{700} \times 100$$

$$= \frac{11}{10} \times 42{,}86 + \frac{20}{20} \times 42{,}86 + \frac{1{,}3}{1} \times 14{,}29$$

$$= 108{,}58$$

O sea

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{\text{precio frutas y verduras ago08}}{\text{precio frutas y verduras jul08}} \times 42{,}86 + \frac{\text{precio carnes ago08}}{\text{precio carnes jul08}} \times 42{,}86 + \frac{\text{precio boleto de colectivo ago08}}{\text{precio boleto de colectivo jul08}} \times 14{,}29$$

$$= \frac{11}{10} \times 42{,}86 + \frac{20}{20} \times 42{,}86 + \frac{1{,}3}{1} \times 14{,}29$$

$$= 108{,}58$$

Expresión general para el cálculo del índice de precios, en el período t, con base en el período 0.

Con esta forma de cálculo el valor del índice de precios del mes de agosto resulta 108,58. El valor no coincide exactamente con el cálculo (1) anterior (108,57), solamente en la segunda cifra decimal, por errores de redondeo.

Expresión general para el cálculo del índice de precios, en el período t, con base en el período 0.

$$\text{número índice} = \sum_{i=1}^{n} \frac{p_t^i}{p_0^i} w^i, \quad \textbf{con} \quad w^i = \frac{p_0^i q_0^i}{\sum_{j=1}^{n} p_0^j q_0^j}$$

p_0^i = precio del producto i en el período 0
p_t^i = precio del producto i en el período t
q_0^i = cantidad del producto i en el período 0
w^i = ponderación del producto i
n = cantidad total de productos que componen la canasta

El índice mide la variación de los precios $\frac{p_t^i}{p_0^i}$ de los productos de la canasta familiar, ponderados por la participación de cada uno de los productos en el valor total de la misma en el período base.

$\left(\frac{p_0^i q_0^i}{\sum_{j=1}^{n} p_0^j q_0^j} \right)$

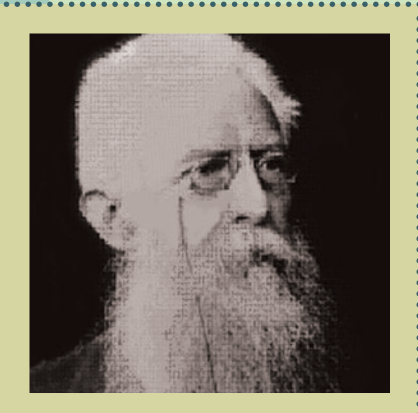
El símbolo $\sum$ indica la suma sobre todos los productos de la canasta (ver sección 18.1.1 para más detalles sobre el uso de $\sum$).

Las ponderaciones utilizadas en el cálculo del índice son 42,86 tanto para "verduras y frutas" como para "carne", y 14,29 para el "boleto de colectivo" (tabla 15.5). ¿Qué significan estas ponderaciones? El costo de la canasta en el período base es de $ 700 (tabla 15.3); "frutas y verduras" con un costo de $ 300 contribuye con un 42,86% del total de la canasta. Lo mismo ocurre con "carne" mientras el gasto por "boleto de colectivo" es un 14,29% del total de la canasta. ¿Por qué decimos que las ponderaciones son fijas? Porque una vez determinada la canasta, sus cantidades y sus precios en el período base (en nuestro ejemplo es julio de 2008) todos los meses se calculará el índice utilizando las mismas ponderaciones.

Cálculo de las ponderaciones de la canasta familiar. TABLA 15.5

Bienes y servicios	Cantidad (julio-08)	Precio por unidad (julio-08)	Costo (jul-08)	Ponderación (%)
Verduras y frutas	30 kg	$ 11/kg	$ 300	(300/700) x 100 = 42,86
Carne	15 kg	$ 20/kg	$ 300	(300/700) x 100 = 42,86
Boleto de colectivo	100 viajes	$1/viaje	$ 100	(100/700) x 100 = 14,29
Costo total			$ 700	100,01

Las ponderaciones de la tabla 15.5 no suman 100 debido a errores de redondeo.



Este tipo de índice, con la canasta familiar fija en sus componentes y cantidades, se denomina **índice de Laspeyres.**

Los índices de precios al consumidor en muchos países se elaboran utilizando el índice de Laspeyres.

A pesar de su nombre francés Ernst Louis Étienne Laspeyres fue un economista y estadístico alemán.

En el cálculo del índice de precios al consumidor **los bienes y servicios se mantienen fijos,** tanto en tipo como en cantidades. Estos bienes y servicios fijos son llamados **canasta familiar.**

El índice de precios al consumidor, IPC, es un número índice para el costo de un **conjunto de bienes y servicios fijo.**

¿Por qué es importante una canasta familiar fija?

Porque de esa forma la comparación es válida. Las diferencias, se deberán únicamente a la variación de los precios. Si la canasta familiar no fuera fija, no podríamos saber si un aumento en el índice se debe a un aumento de los precios, a un aumento de las cantidades consumidas, o a cambios en los productos que se consumen.

¿Cómo obtenemos una canasta familiar para representar a muchas familias? Utilizamos una canasta familiar promedio. Pero, ¡una canasta promedio representa a muchas familias y a ninguna en particular!

Veamos cómo se realizaría el cálculo de la canasta familiar para 2 familias en el mes 1.

Bienes y servicios	Familia 1	Familia 2	Promedio
Verduras y frutas	40 kg	20 kg	30 kg
Carne	0 kg	30 kg	15 kg
Boleto colectivo	80 viajes	120 viajes	100 viajes

La primera familia es vegetariana. La segunda come menos frutas y verduras, pero utiliza más viajes de colectivo que la primera. Si promediamos las cantidades para cada rubro obtendremos las cantidades que presentamos inicialmente (tablas 15.3 y 15.4).

Los institutos de estadística realizan encuestas de hogares a muchas familias para obtener una "canasta familiar promedio". Seguramente no representará a ninguna familia en particular, pero permite evaluar las modificaciones globales de los precios. Los bienes y servicios encuestados se dividen en Rubros.

En el cálculo del Índice de Precios al Consumidor (IPC), con base en el año 1999, el INDEC utilizaba los siguientes rubros:

- Alimentos y bebidas.
- Indumentaria.

- Vivienda.
- Equipamiento y mantenimiento del hogar.
- Atención médica y gastos para la salud.
- Transporte y comunicaciones.
- Esparcimiento.
- Educación.
- Bienes y Servicios Varios.

Todos ellos con una composición específica y fija.

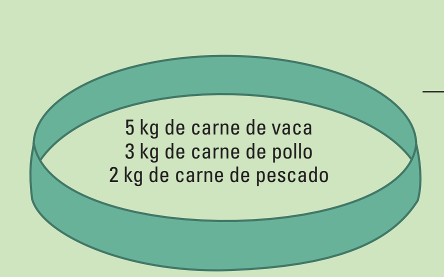
Veamos cómo se relaciona el índice de precios, indicativo de la evolución de la economía en general, con el costo de vida individual.

Consideremos un ejemplo sencillo para ilustrar la diferencia entre el índice de precios al consumidor y el costo de vida.

> El **índice de precios al consumidor** y el **costo de vida** son dos conceptos **distintos.**

Pensemos que Juan consume 10 kg de carne por mes entre carne de vaca, pollo y pescado. El precio por kg en noviembre de cada una de estas carnes es $ 25, $ 16 y $ 20 respectivamente. Por lo tanto, Juan gasta en carnes durante noviembre $ 213, distribuidos de la siguiente manera:

5 kg de carne de vaca	x $25 /kg =	$125
3 kg de carne de pollo	x $16 /kg =	$48
2 kg de carne de pescado	x $20 /kg =	$40
10 kg de carne con un gasto de		$213



Si en diciembre aumenta solamente la carne de vaca, de $25 a $30, el valor de la canasta cárnica será $238:

5 kg de carne de vaca	x 30 $/kg =	$150
3 kg de carne de pollo	x 16 $/kg =	$48
2 kg de carne de pescado	x 2 $/kg =	$40
10 kg de carne con un gasto de		$238

Por lo tanto el valor de la canasta cárnica de diciembre, para el cálculo del índice es $238.

Noviembre $213 Diciembre $238

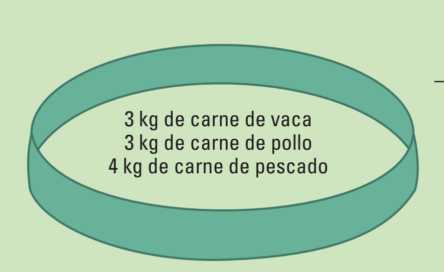
Para obtener la variación del índice se compara el valor de la canasta en el período actual con el anterior:

$$\left(\frac{238}{213}\right) \times 100 = 111{,}74$$

El aumento del IPC será del 11.74 %.

Si los precios no cambiaran Juan seguiría comprando las mismas cantidades; pero, si alguno de los precios aumenta puede **decidir cambiar.** Para un consumo total de 10 kg de carnes, podría decidir reducir su consumo de vacuna y aumentar el pescado, manteniendo su consumo total de carnes y gastar $218:

3 kg de carne de vaca	x 30	$/kg =	$90
3 kg de carne de pollo	x 16	$/kg =	$48
4 kg de carne de pescado	x 20	$/kg =	$80
10 kg de carnes con un gasto de			$218



Juan cambió la canasta.
El índice no.

Con este gasto Juan considera que mantiene su nivel de vida y su costo de vida pasó de

Noviembre $213 Diciembre $218

En porcentaje, ese aumento es de sólo 2,3 %:

$$\left(\frac{218 - 213}{218}\right) \times 100 = 2{,}3$$

Un número índice permite resumir los valores de muchos ítems, para seguir su evolución en el tiempo. Mientras los bienes y servicios de la canasta representen adecuadamente los hábitos de la población, se mantiene fija. Con el tiempo, la canasta tiende a desactualizarse, y se requieren sucesivas adaptaciones para hacerla representativa de una realidad cambiante.

□ 15.5. Mediciones precisas y exactas

Si, por ejemplo, utilizamos la balanza de una farmacia para medir el peso dará una medida válida. Esa balanza, como ocurre a veces, puede no ser muy exacta. Si la balanza mide siempre 1 kg de más valdrá:

peso medido = peso verdadero + 1 kg

Además si repetimos la medición no obtendremos el mismo valor; la balanza no es precisa.

A veces el resultado será un poco mayor:

peso medido = peso verdadero + 1 kg + 0,25 kg

y otras un poco menor:

peso medido = peso verdadero + 1 kg - 0,75 kg

Tenemos dos tipos de errores:

Cuando las mediciones no tienen sesgo decimos que son exactas, y cuando el error aleatorio, que nunca se puede eliminar, es pequeño se trata de mediciones precisas. Ambos tipos de errores suelen estar presentes en los procesos de medición, y los podemos expresar como:

Modelo de medición: valor medido = valor verdadero + sesgo + error aleatorio

Resumiendo, un proceso de medición tiene:

- Error aleatorio, si mediciones realizadas sobre un mismo objeto dan resultados diferentes.
- Sesgo, si sistemáticamente sobreestima o subestima la propiedad que mide.

Precisión no significa validez. Por ejemplo, el volumen del cerebro no es una medida válida de la inteligencia. Sin embargo, en el siglo 19 Paul Broca sostuvo que sí lo era y propuso un método muy preciso para calcularlo (http://www.comoves.unam.mx/articulos/cerebro.shtml). Una de las consecuencias de esta idea era pensar que las mujeres son menos inteligentes que los hombres, porque sus cerebros son más pequeños (como también el resto del cuerpo), coincidiendo con los prejuicios de la época. Actualmente, no existen indicios de diferencias intelectuales entre géneros y se sabe que el volumen cerebral no tiene relación con la inteligencia.

Ningún proceso de medición es perfectamente preciso. **Los promedios mejoran la precisión.** Generalmente, los resultados de los análisis clínicos son el promedio de tres o más mediciones repetidas. Incluso, en las escuelas, los alumnos realizan varias mediciones en sus clases de laboratorio y las promedian.

Reducir el sesgo no es tan fácil porque proviene de la calidad del instrumento. En este caso, es necesario **mejorar el método de medición** para **no tener sesgo.**

¿Qué relación tiene el sesgo aquí descripto con el de la sección 6.3? ¡Se trata del mismo concepto! Dijimos en ese caso: "es un favoritismo de alguna etapa del proceso de recolección de datos". Ese favoritismo producirá una subestimación o una sobrestimación sistemática de la característica de la población a medir. La diferencia fundamental se encuentra en la interpretación del término del error. Cuando se seleccionan individuos de una población y se observa el valor de una variable, el término llamado "error aleatorio" representa las diferencias entre el valor individual (si se seleccionara sin sesgo) y la media poblacional (μ). El modelo general será:

valor individual = μ + sesgo + error aleatorio

Si los individuos se seleccionan mediante un **muestreo aleatorio** simple, eliminamos el sesgo y el modelo resulta:

valor individual = μ + error aleatorio

Cuando se trabaja con datos es importante preguntarse: ¿Cómo se obtuvieron esos números? Si se trata de mediciones sobre muchos individuos, tendremos valores de **variables** describiendo a cada uno de ellos. Debemos saber cómo está definida exactamente cada variable y si se trata de **variables válidas** como mediciones numéricas de los conceptos en estudio.

También es necesario conocer si los datos tienen **errores de medición** que puedan reducir su utilidad. Algunos procedimientos de medición pueden introducir sesgo, en ese caso es necesario **utilizar un instrumento mejor.** Si medir al mismo individuo produce resultados diferentes, de manera que los valores no son confiables, se puede mejorar la confiabilidad **repitiendo la medición varias veces** y utilizando su promedio.

15.6 Actividades y ejercicios

1. Considerando "la inteligencia" como la capacidad de resolver problemas en general, explique por qué no es válido medir la inteligencia preguntando:

   ¿Quién escribió el Martín Fierro?
   ¿Quien ganó el último mundial de fútbol?

2. Un estudio en una ciudad muestra un promedio de 3 muertes de chicos por año en accidentes con micros colectivos y un promedio de 20 muertes en accidentes con autos particulares durante el horario escolar. Estos datos sugieren que viajar en colectivo es más seguro que viajar en auto con los padres. Sin embargo, estas cifras no cuentan toda la historia. ¿Qué comparaciones deberían hacerse para evaluar la seguridad de los dos medios de transporte?

3. El Ministerio de Salud le interesa conocer el progreso alcanzado en la lucha contra el cáncer. Algunas de las variables:

   a) Cantidad total de muertes por cáncer.
   b) Porcentaje de muertes por cáncer.
   c) Porcentaje de pacientes vivos 5 años después del diagnóstico de la enfermedad.

   Ninguna de las variables anteriores es una medida totalmente válida de la efectividad de los tratamientos. Explique cómo a) y b) podrían disminuir y c) aumentar, incluso cuando los tratamientos no fueran efectivos.

4. Interesa estudiar el "estado físico" de las alumnas de 5to año de una escuela. Describa una manera claramente inválida de medir "estado físico". Luego describa un proceso le parezca válido.

16. Variables numéricas

□ 16.1. Histogramas y distribuciones de frecuencias

La **distribución** de una variable nos dice **cuáles son los valores** que puede tomar y su **frecuencia,** es decir, cuántas veces ocurre cada uno de los valores.

Como hemos visto, las tablas de frecuencias y los gráficos (circulares, de barras) permiten conocer la distribución (ya sea en una población o en una muestra) de los valores de una variable categórica. La distribución de los valores de la variable dentro de las diferentes categorías se puede expresar en cantidades, en proporciones o en porcentajes.

Para representar gráficamente la distribución de los datos correspondientes a una **variable numérica** (discreta o continua) también se utilizan tablas de frecuencias y un gráfico similar al gráfico de barras: el histograma.

Un **histograma** representa, en el eje horizontal, **los valores de una variable numérica** divididos en **intervalos de clase.** En forma similar a los gráficos de barras, tiene una barra sobre cada intervalo cuya **altura indica la cantidad** (frecuencia) o **proporción** (frecuencia relativa) de datos. No se deja espacio entre las barras ó rectángulos.

Cuando los valores posibles de la **variable numérica** son pocos, la altura de cada rectángulo del histograma muestra directamente la cantidad o proporción de veces que **cada uno de los valores** ocurrió. Cuando son muchos, es necesario agruparlos definiendo previamente los intervalos.

16.1.1. Variables discretas

Una variable numérica es **discreta** cuando únicamente puede tomar valores dentro de una sucesión determinada de números. La cantidad de hermanos por alumno de una escuela es una variable discreta: puede tomar los valores 0, 1, 2, 3, 4, pero nunca valores como 2,50; 7,2; 0,30.

Veremos primero un ejemplo de una variable numérica discreta (cantidad de hijos) con **pocos valores posibles. No es necesario agruparlos.**

Ejemplo 16.1. Supongamos que se entrevistan 1.000 familias de la Ciudad de Buenos Aires, para saber cuántos hijos tiene cada familia. Nuestros datos son de la forma 0, 0, 3, 1, 1, 1, 2, 2, 2, 3, 1, 1, 2, 0, 0, 0, 2, 1, 8, 1, 1, 2, 3, 0, 0, 0...

Cada número es la cantidad de hijos de cada una de las familias entrevistadas. Es necesario resumir la información: 250 familias no tienen hijos, 200 tienen 1 hijo, 300 tienen 2 hijos, 160 tienen 3 hijos, 50 tienen 4 hijos, 20 tienen 5 hijos, 10 tienen 6 hijos, 7 tienen 7 hijos, 2 familias tienen 8 hijos y una familia tiene 9 hijos. Podemos presentar el resumen mediante la siguiente tabla de frecuencias:

Cantidad de hijos	Frecuencia
0	250
1	200
2	300
3	160
4	50
5	20
6	10
7	7
8	2
9	1
Total	1.000

Tendremos una visualización más rápida de los datos si los representamos mediante un histograma.

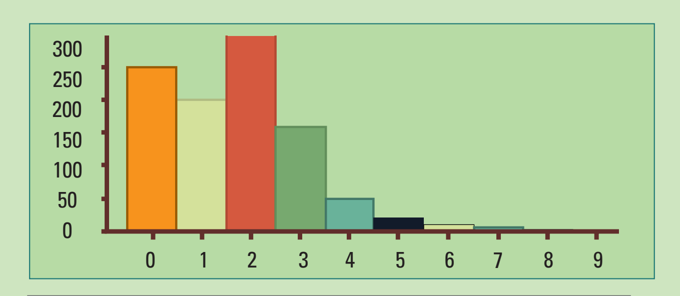


Figura 16.1. *Histograma de la cantidad de hijos por familia, expresado en frecuencias.*

La mayor cantidad de familias tienen 2 hijos, le siguen las familias sin hijos y después las de un sólo hijo.

> Un histograma representa **la distribución de una variable** numérica **en una población** o **en una muestra.** Los intervalos de clase de una variable discreta están centrados en sus valores posibles y tienen la misma longitud.

En el ejemplo 16.1 los datos corresponden a una muestra de 1.000 familias de la Ciudad de Buenos Aires.

¿Cuál es la variable numérica y cuál es la población? ¿Cuáles son los valores posibles de esa variable numérica en la población? ¿Cuál es el tamaño de la muestra?:

- Variable numérica discreta: cantidad de hijos por familia.
- Valores posibles: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
- Población: todas las familias de la Ciudad de Buenos Aires, en un año fijo.
- Tamaño de la muestra: 1.000

Si la muestra es representativa de las familias de la Ciudad de Buenos Aires en ese momento, podremos considerar al histograma, una estimación de la distribución de la variable cantidad de hijos por familia en la población. ¡Un verdadero trabalenguas!

Cuando interesa comparar la frecuencia entre categorías, como ocurre con los diagramas de barras, puede ser más interesante que el eje vertical esté expresado en frecuencias

relativas (es decir proporciones). Por ejemplo, si queremos estudiar el comportamiento social respecto a la cantidad de hijos, saber que el 75% de las familias tienen como máximo dos hijos es más informativo que saber que son 750.

Cantidad de hijos	Frecuencia	Frecuencia relativa	Porcentaje
0	250	250/1.000 = 0,250	25,0
1	200	200/1.000 = 0,200	20,0
2	300	300/1.000 = 0,300	30,0
3	160	160/1.000 = 0,160	16,0
4	50	50/1.000 = 0,050	5,0
5	20	20/1.000 = 0,020	2,0
6	10	10/1.000 = 0,010	1,0
7	7	7/1.000 = 0,007	0,7
8	2	2/1.000 = 0,002	0,2
9	1	1/1.000 = 0,001	0,1
Total	1000	1 1	100,0

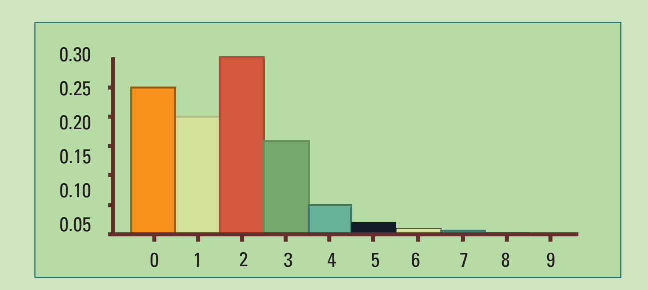


Figura 16.2. *Histograma de la cantidad de hijos por familia, expresado en frecuencias relativas.*

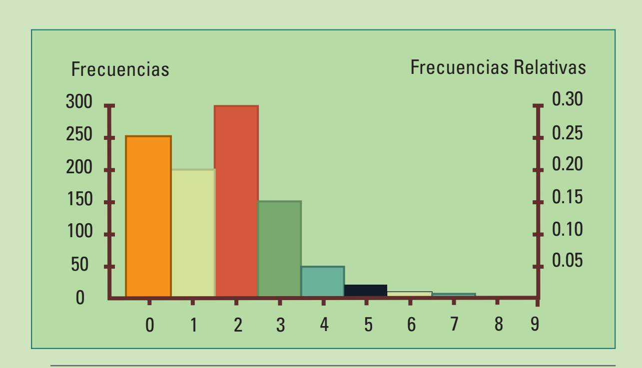


Figura 16.3. *Histograma de la cantidad de hijos por familia, con dos escalas: Frecuencias y frecuencias relativas.*

Observación. Los histogramas de frecuencias y de frecuencias relativas tienen siempre la misma forma, tal como se puede apreciar en las figuras 16.1 y 16.2. Cambian únicamente las escalas verticales. Algunas veces se presentan ambas en el mismo gráfico.

El ejemplo 16.1 (cantidad de hijos por familia) es hipotético. Como es difícil definir “familia”, resulta más realista considerar la cantidad de hijos por mujer, como veremos en el siguiente ejemplo con datos reales.

Ejemplo 16.2. Se trata de la cantidad de hijos de mujeres con edades entre 30 y 34 años en el año 1991 en la Ciudad de Buenos Aires (tabla 16.1); 25.729 mujeres no tienen hijos (24,5%), 19.573 mujeres tienen un solo hijo (18,6%), 33.060 mujeres tienen 2 hijos (31,4%), etc.

El ejemplo 16.1 (cantidad de hijos por familia) es hipotético. Como es difícil definir "familia", resulta más realista considerar la cantidad de hijos por mujer, como veremos en el siguiente ejemplo con datos reales.

Ejemplo 16.2. Se trata de la cantidad de hijos de mujeres con edades entre 30 y 34 años en el año 1991 en la Ciudad de Buenos Aires (tabla 16.1); 25.729 mujeres no tienen hijos (24,5%), 19.573 mujeres tienen un solo hijo (18,6%), 33.060 mujeres tienen 2 hijos (31,4%), etc.

Cantidad de hijos de mujeres, con edades desde 30 a 34 años de la Ciudad de Buenos Aires. Año 1991. TABLA 16.1

Cantidad de hijos	Frecuencia	Frecuencia relativa	Porcentaje
0	25.729	25.729/105.210 = 0,245	24,5
1	19.573	19.573/105.210 = 0,186	18,6
2	33.060	33.060/105.210 = 0,314	31,4
3	18.020	18.020/105.210 = 0,171	17,1
4	5.467	5.467/105.210 = 0,052	5,2
5	1.867	1.867/105.210 = 0,018	1,8
6	813	813/105.210 = 0,008	0,8
7	380	380/105.210 = 0,004	0,4
8	216	216/105.210 = 0,002	0,2
9	85	85/105.210 = 0,001	0,1
Total	105.210	1 1	100,0

Fuente: *Dirección General de Estadística y Censos (G.C.B.A.) sobre la base de datos del Censo Nacional de Población y Vivienda, 1991 - Serie C.*

Un histograma de los datos de la tabla 16.1 nos permite visualizar más rápidamente su distribución.

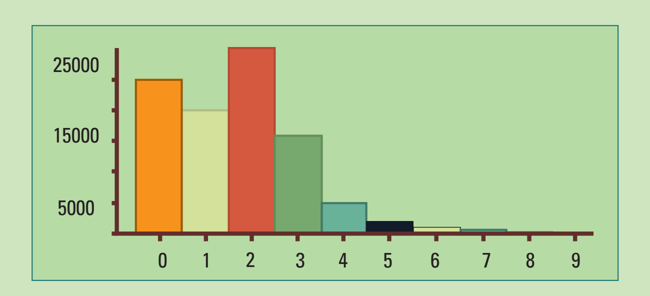


Figura 16.4. *Datos reales. Ciudad de Buenos Aires año 1991. Histograma de la cantidad de hijos por mujer con edades entre 30 y 34 años.*

La frecuencia (escala vertical del histograma, figura 16.4) es la cantidad de mujeres con edades entre 30 y 34 años en el año 1991, con 0,1, 2, ..., hasta 9 hijos, respectivamente en cada intervalo. Se destaca el rectángulo centrado en 2, porque tiene la mayor altura; 2 es la cantidad más frecuente de hijos en la Ciudad de Buenos Aires.

La distribución, es muy parecida a la del ejemplo hipotético; ambos histogramas tienen casi la misma forma pero las frecuencias, frecuencias relativas y porcentajes ya no son números redondos.

¿Cuál es la variable numérica y cuál es la población? ¿Cuáles son los valores posibles de esa variable numérica en la población?:

- Variable numérica discreta: cantidad de hijos por mujer
- Valores posibles: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 (no es posible tener 2,75 hijos).
- Población: todas las mujeres de la Ciudad de Buenos Aires entre 30 y 34 años en el 1991.

¿Puede haber mujeres con más de 9 hijos? Efectivamente, puede haber mujeres con 10 ó más hijos. En la ciudad de Buenos Aires sólo se incluye una categoría de 10 o más, porque son pocas. Para poder comparar las categorías mediante un histograma es necesario que tengan el mismo tamaño; es decir, que correspondan a la misma cantidad de valores posibles de la variable. Por esta razón no se incluyó en el histograma la categoría 10 ó más, correspondiente a los valores 10, 11, 12, 13, 14, etc.

16.1.2. Variables continuas

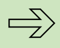
Una variable numérica es continua cuando, dados dos valores posibles de la variable, ésta siempre puede tomar cualquier valor intermedio. El peso de una persona es una variable numérica continua, puede tomar valores como 48 kg ó 49 kg y también, 48,5 kg 48,52 kg etc.

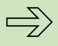
Podemos preguntarnos: ¿cambió la edad a la cual las mujeres tienen hijos? Veamos un ejemplo real para intentar responder esta pregunta. Como la variable edad tiene muchísimos valores posibles, para construir un histograma, los agruparemos en intervalos.

Ejemplo 16.3 Comparemos como se distribuye la edad de las mujeres en el momento del nacimiento de un hijo, en los años 2001, 2003, 2006, utilizando la información del Ministerio de Salud.

NACIMIENTOS EN LA REPÚBLICA ARGENTINA SEGÚN EDAD DE LA MADRE. TABLA 16.2

Año	2001	2003	2006	2001	2003	2006
Grupo de edad	Cantidad			Porcentaje		
[10-15)	3.022	2.763	2.766	0,44	0,40	0,40
[15-20)	97.060	92.461	103.885	14,20	13,25	14,92





Año	2001	2003	2006	2001	2003	2006
Grupo de edad	Cantidad			Porcentaje		
[20-25)	188.415	184.155	174.342	**27,57**	26,39	**25,03**
[25-30)	170.748	179.107	176.931	24,98	25,66	25,40
[30-35)	128.521	137.359	139.003	**18,80**	19,68	**19,96**
[35-40)	68.162	71.497	73.177	**9,97**	10,24	**10,51**
[40-45)	19.658	20.674	19.866	2,88	2,96	2,85
[45-50)	1.417	1.438	1.405	0,21	0,21	0,20
[50-55)	98	92	83	0,01	0,01	0,01
Sin información	6.394	8.406	4.993	0,94	1,20	0,72
Total	683.495	697.952	696.451	100,00	100,00	100,00

Fuente: *Estadísticas Vitales. Ministerio de Salud. 2001, 2003, 2006. ISSN 1668-9054.*

¿Cómo se interpretan los grupos de edad?

El grupo [10-15) corresponde a las edades entre 10 y 15 años
El grupo [15-20) corresponde a las edades entre 15 y 20 años
El grupo [20-25) corresponde a las edades entre 20 y 25 años
El grupo [25-30) corresponde a las edades entre 25 y 30 años
El grupo [30-35) corresponde a las edades entre 30 y 35 años
....
Una edad de 15 años se cuenta en el grupo [15-20) y no en el [10-15)
Una edad de 20 años se cuenta en el grupo [20-25) y no en el [15-20)
....

El intervalo [15-20) es un intervalo cerrado en 15 (se incluye el valor 15 en el intervalo) y abierto en 20 (no se incluye el valor 20 en el intervalo).

¿Cuál es la variable numérica y cuál es la población?:

En general, el intervalo **[a-b),** donde **a** y **b** son números reales cualesquiera con **a** menor que **b**, es un **intervalo cerrado en a** (incluye el valor **a**) y **abierto en b** (**no** incluye el valor **b**)

- Variable numérica continua: edad de la madre en el momento del parto. Es posible tener una edad decimal de 18,75 años (18 años y 9 meses).
- Valores posibles: desde 10 hasta 54 años.
- Población: se consideran en este ejemplo tres poblaciones:
- Todos los niños nacidos en el año 2006.
- Todos los niños nacidos en el año 2003.
- Todos los niños nacidos en el año 2001.

Los histogramas de la figura 16.5 permiten comparar cómo se distribuyen las edades de las madres de la República Argentina en la población de los niños nacidos en el año 2006, 2003 y 2001 respectivamente.

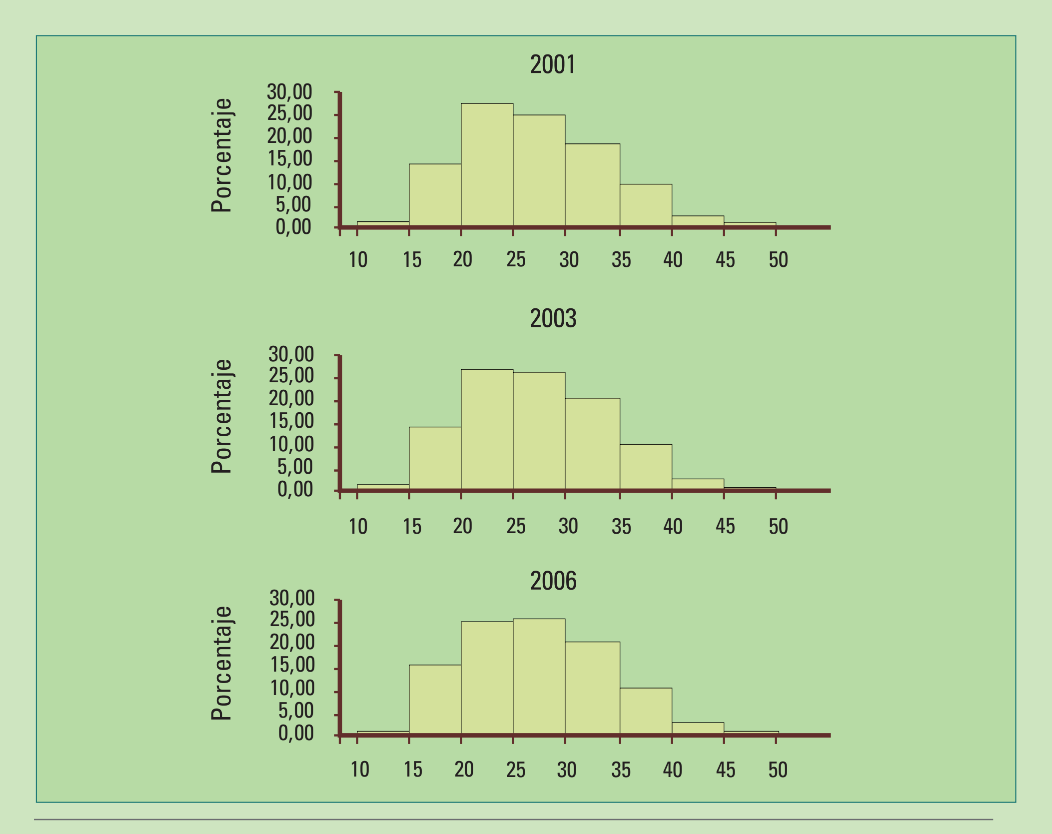
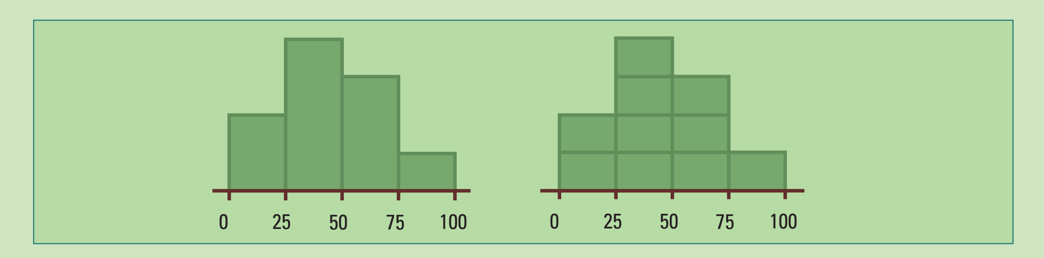


Figura 16.5. *Edad de la madre en momento del parto para los años 2001, 2003, 2006 en la Ciudad Autónoma de Buenos Aires.*

Los 3 histogramas de la figura 16.5 tienen formas similares, esto indicaría que la respuesta a la pregunta planteada es no. No cambiaron las edades en las cuales las mujeres tienen hijos en la República Argentina entre los años 2001, 2003 y 2006. Sin embargo, si observamos con más detalle vemos un porcentaje mayor en el año 2001 de nacimientos provenientes de madres con edades en el intervalo [20-25) años. En el 2003 esa diferencia entre los intervalos [20-25) y [25-30) se hace casi imperceptible y en el 2006 comienza ya el [25-30) tiene un porcentaje de 25,40 % un poco mayor que el del [20-25) con 25,03%. Mirando la tabla 16.2 (pág. 100) podemos ver además, porcentajes crecientes desde el 2001 al 2006 en los grupos de edades [30-35) y [35-40) desde el 2001 al 2006. Estas tendencias favorecen la idea que las mujeres tienen sus hijos a edades cada vez más tardías aunque se mantiene alto, cercano al 15%, el porcentaje de madres adolescentes. Esto es una preocupación de las autoridades sanitarias. La incidencia de prematuros, bajo peso al nacer y de parto instrumentado, es mayor entre las madres adolescentes que en madres con edades entre 20 y 30 años.

En un histograma puede faltar el eje vertical.

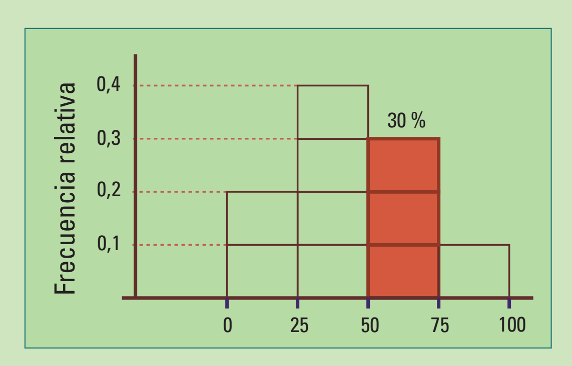
Ejemplo 16.4. Al siguiente histograma le falta el eje vertical. ¿Qué información puede proveer?



Sin el eje vertical no se pueden hallar las frecuencias absolutas, pero, sí es posible determinar las frecuencias relativas de cada uno de los intervalos. Debemos ver qué proporción del área total del histograma se encuentra por encima de cada intervalo. Dividimos la superficie del histograma en 10 rectángulos iguales de manera que cada porción es 1/10 de esa superficie, es decir el 10%.

Hay 2 rectángulos sobre el intervalo 0-25, tienen el 20% del área; 4 rectángulos sobre 25-50, 40% del área; **3 rectángulos sobre 50-75, 30%** y 10% está sobre 75-100.

Generalmente, no es tan fácil dividir a los histogramas en 10 partes iguales, sin embargo siempre las frecuencias relativas se corresponden con áreas relativas.



□ 16.2. Construcción de tablas de frecuencias

En los ejemplos anteriores los datos ya estaban agrupados o los histogramas estaban construidos. Vimos tablas con distribuciones de frecuencias para variables numéricas discretas (ejemplo 16.1 y 16.2) y para una variable numérica continua (ejemplo 16.3).

En las siguientes secciones veremos cómo se agrupan los datos numéricos y se construyen las tablas de frecuencias para obtener los histogramas. Trataremos en forma separada a los datos de variables discretas y continuas.

16.2.1. Variable discreta

Paso 1. Se ordenan los valores posibles de la variable.
Paso 2. Se cuenta cuántas veces aparece un dato con cada valor posible. Esto nos da la frecuencia.
Paso 3. Se divide cada frecuencia por el total de datos, obteniendo así la frecuencia relativa.

Ahora su turno: Registre cuántos hermanos tienen cada uno de los alumnos de su división y obtenga una tabla de frecuencias y de frecuencias relativas. ¿Cuál es la variable? ¿Cuáles son sus valores posibles? A partir de la tabla construya el histograma correspondiente. ¿Cuál es la población en estudio?

16.2.2. Variable continua

Paso 1. Se ordenan los datos.
Paso 2. Se definen intervalos de clase con igual longitud, cubriendo el rango de los valores observados.
Paso 3. Se cuentan cuantos datos pertenecen a cada uno de los intervalos. Esto indica la frecuencia.
Paso 4. Se divide cada frecuencia por el total de datos, obteniendo así la frecuencia relativa.

En el ejemplo siguiente veremos cómo construir la tabla de frecuencias para datos de una variable numérica continua.

Ejemplo 16.5. Los datos siguientes corresponden al peso (en kg) de 52 alumnos y 49 alumnas de 3 divisiones de 4to. año.

- **Varones:** 67 57 64 73 65 69 67 66 67 69 63 65 66 53 58 64 69 67 63 71 69 62 59 61 72 68 57 55 79 59 66 58 72 67 71 67 65 61 63 69 74 64 66 70 63 51 79 68 67 66 85 81

- **Mujeres:** 46 52 52 52 51 43 48 44 55 43 50 57 52 54 51 54 48 48 62 52 50 52 45 54 47 50 50 51 60 56 51 52 54 42 54 48 50 56 50 48 52 55 54 58 46 37 38 68 70

¿Cuál es la variable? Peso
¿Es una variable numérica continua o discreta? El peso es una variable numérica continua.

¿Cuál es la población?
Si nos interesa describir el peso de los/as alumnos/as de esas 3 divisiones de 4to. año, la población está formada por todos/as los alumnos/as de esas 3 divisiones.

¿Qué podemos decir de la distribución de los pesos mirando estos datos?
Para comenzar construiremos un diagrama de puntos, donde cada punto corresponde a un alumno de ese peso. Los valores repetidos se ponen uno encima del otro, a iguales distancias. ¿Se puede ver algo raro? Hay espacios vacíos y se distinguen 2 picos.



Figura16.6. *Diagrama de puntos de los pesos de varones y mujeres de 4to. año.*

Luego, construiremos una tabla de frecuencias, para eso se dividimos la recta numérica en intervalos de clase y contamos cuántos pesos caen dentro de esos intervalos. La frecuencia relativa es la proporción de pesos dentro de cada intervalo.

FRECUENCIAS DE LOS PESOS (EN kg) DE LOS ALUMNOS Y ALUMNAS DE 4TO. AÑO. TABLA 16.3

Intervalo de Clase	Frecuencia	Frecuencia relativa
[30 - 45)	6	
[45 - 60)	48	
[60 - 75)	43	
[75 - 90)	4	
		COMPLETAR
Total	101	1

El intervalo [30-45) es un intervalo cerrado en 30 (se incluye el valor 30 en el intervalo) y abierto en 45 (no se incluye el valor 45 en el intervalo).

El intervalo [45-60) es un intervalo cerrado en 45 (se incluye el valor 45 en el intervalo) y abierto en 60 (no se incluye el valor 60 en el intervalo).

El número al lado del corchete se incluye en el intervalo, el número al lado del paréntesis no.

Ahora su turno. Completar:

El intervalo [60-75) es un intervalo cerrado en .... y abierto en ...., porque

....

El intervalo [75 - 90) es un intervalo cerrado en .... y abierto en ...., porque

....

FRECUENCIAS DE LOS PESOS (EN kg) DE LOS ALUMNOS DE 4TO. AÑO. TABLA 16.4

Intervalo de Clase	Frecuencia	Frecuencia relativa
[30 - 35)	0	
[35 - 40)	2	
[40 - 45)	4	
[45 - 50)	9	
[50 - 55)	26	
[55 - 60)	13	
[60 - 65)	12	
[65 - 70)	23	
[70 - 75)	8	
[75 - 80)	2	
[80 - 85)	1	
[85 - 90)	1	
		COMPLETAR
Total	101	1

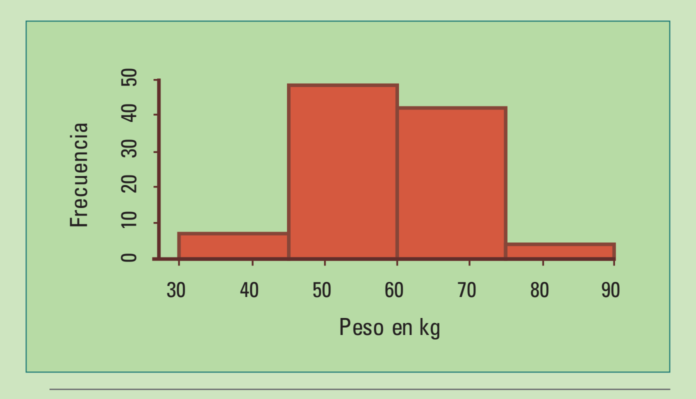


Figura 16.7.

Ahora, se debe construir el histograma. Éste (figura 16.7) no parece demasiado interesante. La mayoría de los pesos se encuentran entre los 45kg y los 75 kg, entonces podemos subdividir los intervalos de clase en tres partes iguales y obtenemos una nueva tabla de frecuencias (tabla 16.4).

El primer intervalo de clase [30-35) no tiene datos, por lo tanto ningún/a alumno/a tiene su peso dentro de ese intervalo. ¿Qué significan el corchete y el paréntesis?

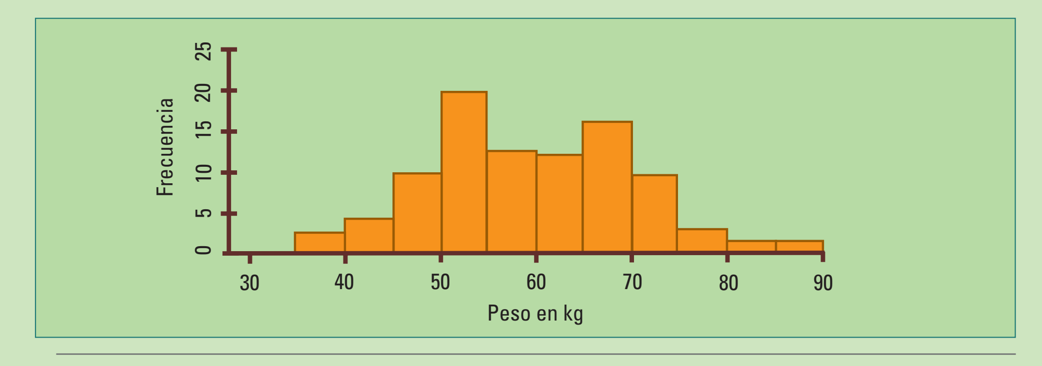


Figura16.8. *Histograma los pesos de varones y mujeres de 4to. año.*

Ahora el histograma (figura 16.8), de manera similar al diagrama de puntos (figura 16.6), nos muestra una información más interesante de la distribución de los pesos. Ambos sugieren la presencia de dos grupos aunque no se vean totalmente separados. En este ejemplo, conocemos los dos grupos mezclados, varones y mujeres. En el histograma se puede apreciar además, el carácter continuo de la variable peso.

No hay una regla para obtener la cantidad más conveniente de intervalos de clase, pero daremos unas ideas al respecto:

- Utilice intervalos de igual longitud centrados en valores redondos, si es posible, enteros.
- Si tiene pocos datos utilice una pequeña cantidad de intervalos.
- Para conjuntos de datos más grandes utilice más cantidad de intervalos.
- Una cantidad adecuada suele ser entre 6 y 12 intervalos.

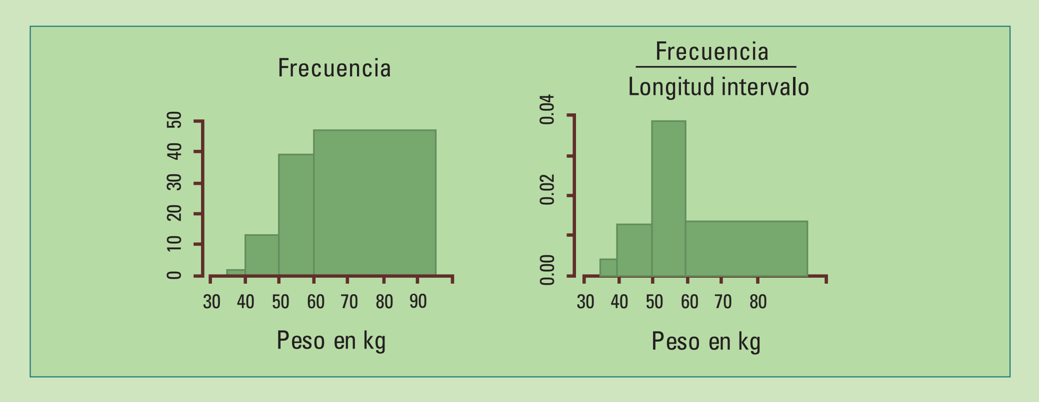
16.2.2.1. Un detalle extra

¿Pueden los **intervalos de clase** de un histograma tener **longitudes diferentes?**

Pueden, pero su construcción se complica.

En ese caso, para la altura del rectángulo de cada clase es necesario utilizar la frecuencia o la frecuencia relativa **dividida** por **la longitud** de dicho **intervalo de clase** (llamada **escala densidad**), de lo contrario, aumentar la longitud implicaría aumentar la altura, y disminuir su longitud resultaría en reducir la altura, **distorsionando** artificialmente la forma del histograma.

La figura siguiente muestra dos histogramas, en el de la izquierda la escala vertical es la frecuencia, y en el de la derecha, la frecuencia relativa dividida la longitud del intervalo de clase.



En el histograma de la izquierda, de **frecuencias absolutas** de los pesos de alumnas y alumnos de 4to. año, utilizando **intervalos de clase de distinta longitud, no representa** adecuadamente **la distribución de los datos** (ver figuras 16.7 y 16.8). Muestra más alumnos entre 60 y 90 kg que entre 30 y 60 kg. El de la derecha mejora la representación de la distribución de los datos.

Conclusión. Siempre que pueda **utilice intervalos de clase de la misma longitud.** Si no es posible elija la escala de densidad para el eje vertical.

◻ 16.3. Diagrama tallo - hoja

Los histogramas son adecuados para conjuntos grandes de datos. Muestran su distribución pero se pierden los valores individuales. Para conjuntos con alrededor de 100 datos o menos, preferimos utilizar un diagrama tallo-hoja pues muestra no sólo la distribución de los datos sino también sus valores.

El estadístico John Tukey propuso en 1975, los diagramas tallo-hoja, una forma rápida para mostrar la distribución de datos correspondientes a variables numéricas, sin necesidad de obtener tablas de frecuencias, conservando todos los valores.

En estos diagramas las filas juegan el mismo papel de los rectángulos de clase en un histograma. Son como un **histograma girado 90º.** Cada fila está encabezada por un número, llamado **tallo,** a continuación se coloca una **línea vertical** y luego **las hojas.** Los valores de los tallos indican en forma compacta los intervalos de clase y tienen valores crecientes hacia abajo. Las hojas representan a los datos.

A continuación, construimos un diagrama tallo-hoja con los datos del ejemplo 16.5, el peso de alumnos y alumnas:

- **Varones:** 67 57 64 73 65 69 67 66 67 69 63 65 66 53 58 64 69 67 63 71 69 62 59 61 72 68 57 55 79 59 66 58 72 67 71 67 65 61 63 69 74 64 66 70 63 51 79 68 67 66 85 81

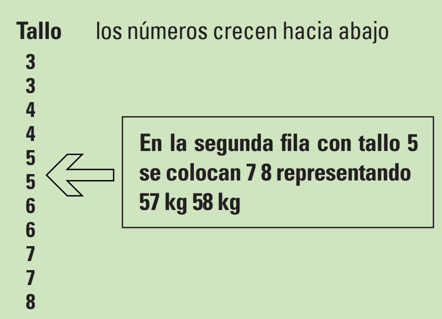
- **Mujeres:** 46 52 52 52 51 43 48 44 55 43 50 57 52 54 51 54 48 48 62 52 50 52 45 54 47 50 50 51 60 56 51 52 54 42 54 48 50 56 50 48 52 55 54 58 46 37 38 68 70

Intervalo	Tallo	Intervalo Tallo	Tallo
[30, 35)	3	[60, 65)	6
[35, 40)	3	[65, 70)	6
[40, 45)	4	[70, 75)	7
[45, 50)	4	[75, 80)	7
[50, 55)	5	[80, 85)	8
[55, 60)	5	[85, 90)	8

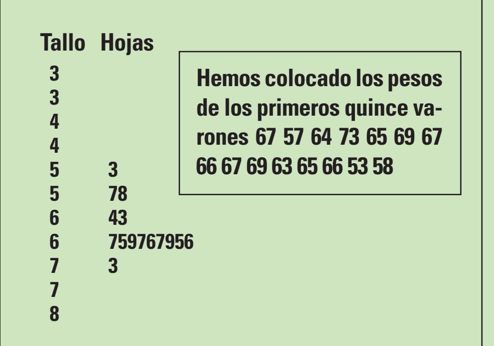
Elegimos los intervalos de clase y les asignamos su tallo

Los tallos están repetidos, aparecerán en el diagrama en dos filas consecutivas. En la fila superior van las hojas desde el cero al 4 y en la inferior las hojas desde el 5 al 9. Por ejemplo, el 5 de la fila superior representa al intervalo [50, 55] y allí se colocan las hojas (el segundo dígito) de todos los datos de ese intervalo y en la inferior se colocan las hojas de todos los datos del intervalo [55, 60].

El **tallo** es una columna de números correspondientes al primer dígito de los datos (dejamos el segundo dígito para las hojas)



Colocamos **el segundo dígito,** la hoja, en la fila adecuada



Ya hemos completado el diagrama con todos los datos

Tallo	Hojas
3	
3	78
4	3432
4	688857886
5	31222102414202400112440024
5	7897598576658
6	43432113320
6	7597679569798675968768
7	3122400
7	99
8	1
8	5

Finalmente ordenamos los valores de las hojas

Tallo	Hojas
3	
3	78
4	2334
4	566788888
5	000000111112222222223444444
5	5556677788899
6	011223333444
6	5556666677777788899999
7	00112234
7	99
8	1
8	5

17. Tipos de distribuciones

Las distribuciones Normales pueden ser las raras.

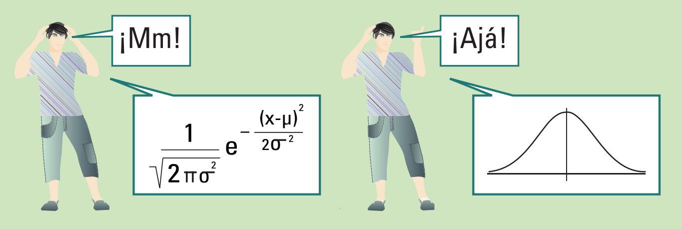
17.1. Distribución Normal

Los histogramas y los diagramas tallo-hoja permiten visualizar cómo se distribuyen los valores de una variable numérica. **Muchas veces estos gráficos tienen la forma de una campana,** con una zona central en la cual los valores de la variable son más frecuentes. A medida que nos alejamos de esa zona central las frecuencias disminuyen simétricamente.



Figura 17.1. *Conjunto de datos con distribución en forma de campana, denominada Distribución Normal*

Esta forma de campana también es llamada **campana de Gauss.** Fue descubierta por Abraham de Moivre en 1720. En 1809, Carl Friedrich Gauss, la utilizó para describir los errores de observación cometidos por los astrónomos, al tomar medidas en forma repetida. Fue denominada **curva de error.**



Su fórmula es:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

En esta expresión matemática, aparecen e (un número aproximadamente igual a 2,718), π (el ya famoso 3,1415), μ (el centro de la campana) y σ (que permite variar su ancho).



Johann Carl Friedrich Gauss (1777–1855), físico y matemático alemán. Fue un niño prodigio.

Cuentan que en la escuela, para que los alumnos se queden tranquilos por un rato, el maestro les dio la tarea de sumar los números del 1 al 100. Inmediatamente Gauss respondió 5.050. Se había dado cuenta que la suma de los extremos, y a medida que avanzaba, siempre daba 101:

1+ 100 = 101
2+ 99 = 101
3+ 98 = 101

....hasta llegar a la mitad, 50. Sumar todos es 50 veces 101, o sea 50 x 101 = 5.050

La campana de Gauss se obtiene graficando los pares (x, f (x)) en el plano:

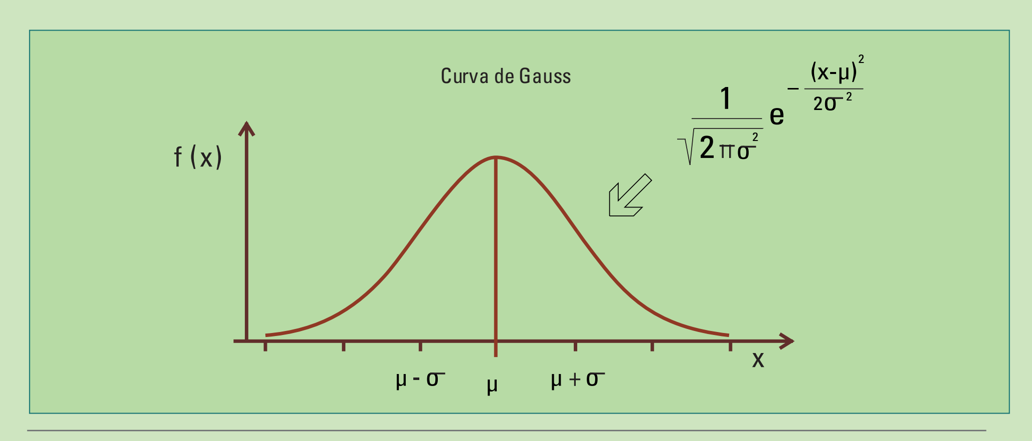


Figura 17.2. *Curva de Gauss junto con su expresión matemática.*

Sólo utilizaremos la forma de la curva y no su expresión.

En 1836 el astrónomo, meteorólogo, estadístico y sociólogo belga Adolphe **Quetelet** extendió la aplicación de la curva, y la utilizó para describir las variaciones de ciertas variables antropomórficas (medidas del cuerpo humano: peso, altura, etc.) entre individuos.

A partir de Quetelet, Francis **Galton** (primo de Charles Darwin y pionero en estudios de genética y de los mecanismos de la herencia) se enteró de la existencia de la curva y **se enamoró de ella.** Dicen que exclamó: "¡Si los griegos la hubieran conocido la habrían deificado!". Galton la llamó curva Normal por primera vez en 1889.

Cuando los datos se distribuyen en forma de campana decimos que tienen distribución Normal o Gaussiana. En la práctica, **los datos rara vez serán "perfectamente Normales"** pero muchas veces la **campana de Gauss** es una muy buena **aproximación al histograma** de un conjunto de datos.

17.1.1. Curva Normal estándar.

Si = 0 y =1 la curva Normal es:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Consideremos un conjunto de datos cuyo histograma puede aproximarse por la curva Normal con parámetros μ y σ . Si a cada dato se le resta μ y se lo divide por σ, entonces el histograma del nuevo conjunto de datos podrá aproximarse por la curva Normal Estándar (figura 17.3).

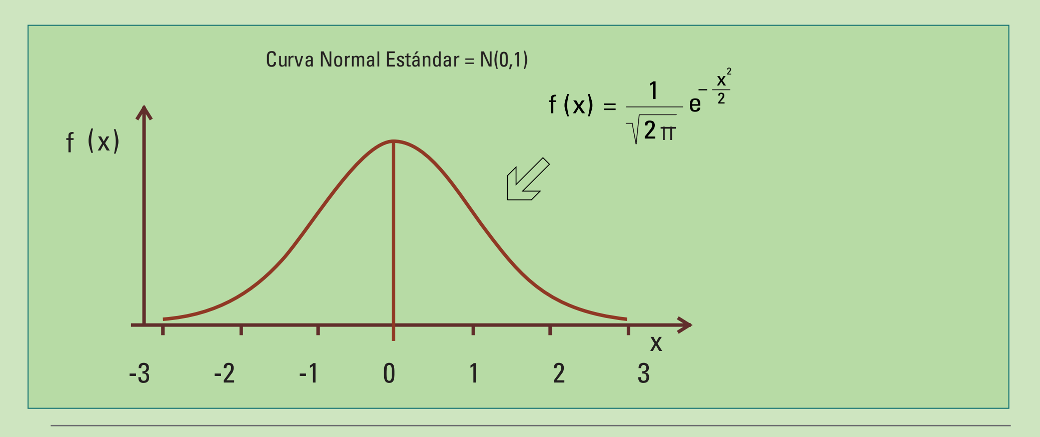


Figura 17.3. *Curva Normal Estándar junto con su expresión matemática.*

17.1.2. ¿Cuándo se obtienen datos con variabilidad Normal y cuando no?

Insistimos, un conjunto de datos rara vez podrá tener una distribución que se ajuste perfectamente a la curva Normal. Sin embargo, en muchas situaciones, esta curva provee una excelente aproximación a los histogramas de los datos. A continuación, presentamos algunos ejemplos en los cuales la aproximación puede ser buena y algunos de cuando no puede serlo.

Ejemplo 17.1. Variabilidad Normal entre unidades muestrales.

Piezas metálicas producidas con la misma máquina, por el mismo operador y en el mismo turno, podrán parecer iguales, pero al medir su **dureza** con cuidado se encontrarán **diferencias.** Cuando estas variaciones se producen en **condiciones normales** (ahora con ene minúscula) – con esto queremos decir: la máquina está funcionando como habitualmente, la materia prima es la de siempre, las herramientas están como todos los días, los operarios descansados y con el ánimo de siempre - entonces las piezas serán parecidas. Las variaciones, respecto de alguna variable (dureza, longitud, peso, elasticidad), darán muchos valores en el centro y pocos en los extremos. A esto lo denominamos "variabilidad Normal" (aquí con ene mayúscula). En cambio, si el producto se fabrica con materias primas defectuosas o los operarios estaban distraídos y siguieron operando cuando la herramienta estaba dañada, la distribución de las variables examinemos ya no será Normal.

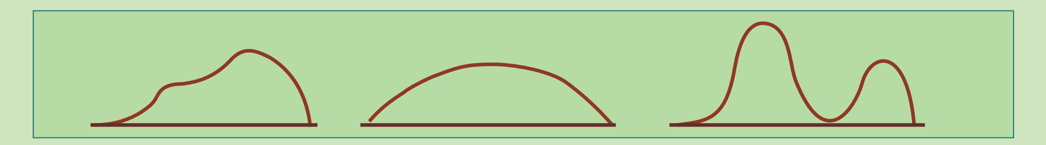


Figura 17.4. *Variabilidad no normal.*

Una aclaración: "normal" con ene minúscula es sinónimo de "habitual"; "Normal" con ene mayúscula se refiere a una distribución de datos en forma de campana

Ejemplo 17.2. Variabilidad debido al error de medición.

Si evaluamos varias veces la **dureza de un mismo** producto, los valores no serán idénticos, aunque la dureza sea siempre la misma. Habrá muchos valores cercanos al **valor verdadero** de la dureza y la frecuencia de los valores disminuirá al alejarse. Si estas mediciones son realizadas con mucho cuidado, por el mismo operador, realizando desde el comienzo los mismos pasos hasta obtener el resultado, su histograma tendrá la forma de la curva Normal. En cambio, si hubo algún descuido al realizar las mediciones, cambió el operador o alguna condición del proceso de medición, estas no tendrán un histograma que pueda ser representado mediante la curva de Gauss.

Ejemplo 17.3. Variabilidad biológica normal.

Si registramos las estaturas de las niñas de una división, encontraremos unas pocas son muy bajitas, otras pocas muy altas y la mayoría con alturas intermedias. Los datos, las mediciones, tendrán una distribución aproximadamente Normal. En cambio, si consideramos las alturas de todos los alumnos (varones y mujeres), los datos provienen de muestras no homogéneas y no tendremos una "variabilidad Normal". Pero no todas las variables antropomórficas tendrán una buena aproximación por la distribución gaussiana como creía Quetelet, por ejemplo, el peso de las personas de una edad y género determinados no tiene distribución simétrica, tampoco los niveles de los triglicéridos en sangre.

La **Normalidad estadística** no implica la normalidad biológica, social o económica. Muchas veces las distribuciones Normales son las raras.

La distribución de los salarios de una población, el caudal de un río de montaña, la precipitación diaria en cierta ciudad, son ejemplos de distribuciones asimétricas.

☐ 17.2. Formas que describen diferentes tipos de distribuciones. Curvas de densidad.

La figura 17.5 muestra un histograma de 400 datos de una variable continua y una curva que describe la forma con la que se distribuyen los datos a lo largo de sus valores. Vemos que las mayores frecuencias se encuentran entre cero y cuatro. Para valores mayores que cuatro se reduce constantemente la frecuencia. Podríamos pensar que la curva se obtiene en dos pasos:

- dibujando el borde superior de cada rectángulo de clase y luego.
- suavizando los escalones.

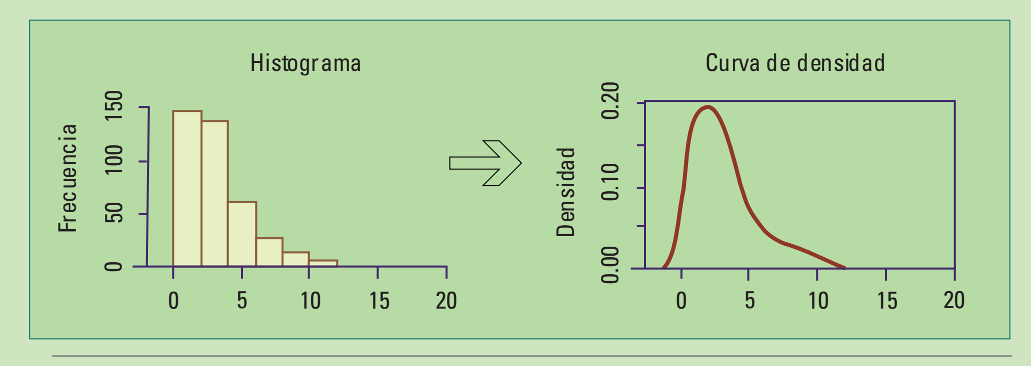
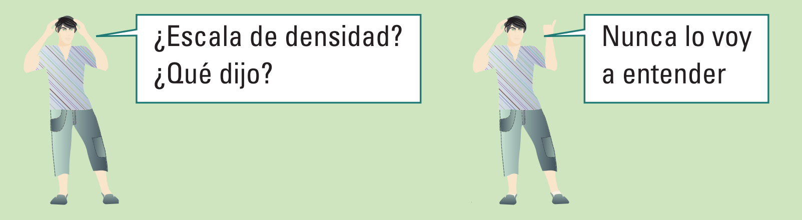


Figura 17.5. *Un histograma y su curva de densidad.*

Existe una diferencia para resaltar entre histogramas y curvas de densidad. Las curvas de densidad se grafican en escala de densidad y los histogramas en escala de frecuencias o frecuencias relativas.

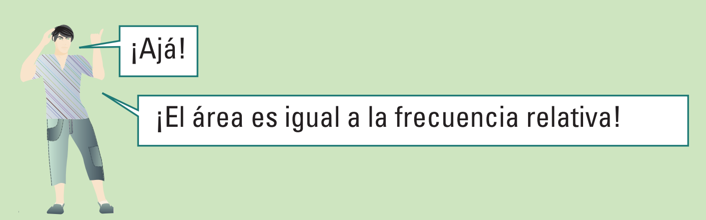


La mayoría de los histogramas muestran la cantidad (frecuencia) o proporción (frecuencia relativa) de observaciones de cada intervalo de clase mediante la altura del rectángulo. De esta manera, el **área de cada rectángulo es proporcional a la frecuencia relativa.** En una **escala de densidad,** el **área de cada rectángulo** es **IGUAL a la frecuencia relativa,** y se obtiene graficando en el eje vertical la frecuencia relativa dividida la longitud del intervalo de clase. En escala de densidad, el área total de los rectángulos del histograma es 1.

Para los datos de la figura 17.5 la longitud de los intervalos es 2 y tenemos:

Escala	Frecuencias		Frecuencias relativas		Densidad	
	Altura	Área	Altura	Área	Altura	Área
	147	294	0,3675	0,735	0,18375	0,3675
	138	276	0,3450	0,690	0,17250	0,3450
	62	124	0,1550	0,100	0,07750	0,1550
	29	58	0,0725	0,145	0,03625	0,0725
	16	32	0,0400	0,080	0,02000	0,0400
	6	12	0,0150	0,030	0,00750	0,0150
	1	2	0,0025	0,005	0,00125	0,0025
	0	0	0,0000	0,000	0,00000	0,0000
	1	2	0,0025	0,005	0,00125	0,0025
Total	400	800	1,0000	2,000	0,5000	1,0000

En **escala de densidad** el **área del rectángulo** de clase es igual a la **frecuencia relativa** y la suma de las áreas es 1.



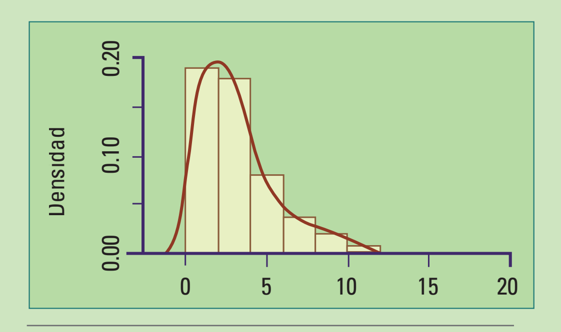
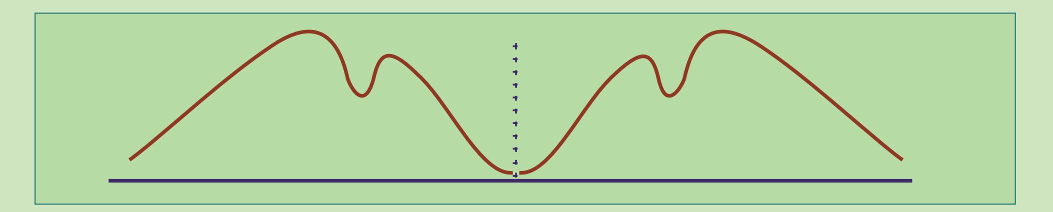


Figura 17.6. *Superposición de un histograma y una curva de densidad.*

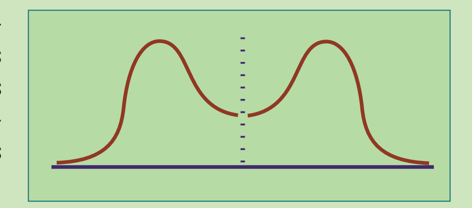
Los histogramas pueden tener distintas formas. Mostraremos algunos patrones especiales en forma simplificada mediante **curvas,** también llamadas **curvas de densidad.** La campana de Gauss es una de ellas.

17.2.1. Simétrica

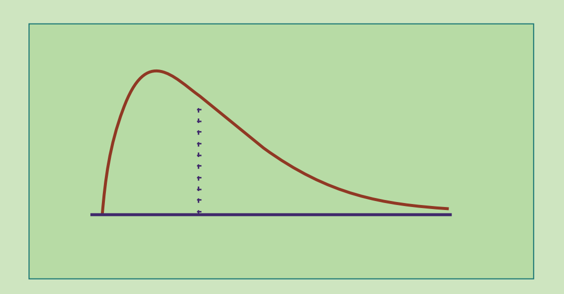
Una **distribución es simétrica** cuando sus dos mitades son imágenes especulares una de la otra.



Por ejemplo, un histograma de las alturas de los mayores de 18 años de un pueblo tendrá dos zonas más altas en espejo, una para los varones y otro para las mujeres, mientras haya la misma cantidad de varones y mujeres. Esto se debe a la superposición de dos curvas simétricas con distinto centro e igual ancho.



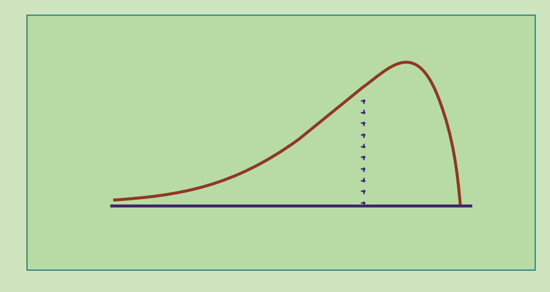
17.2.2. Asimétrica a derecha



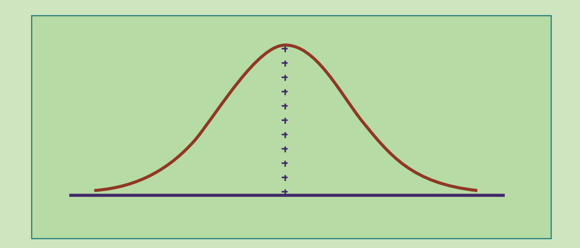
Una **distribución es asimétrica a derecha** cuando la mitad derecha es más finita y más larga. Por ejemplo, la distancia de los domicilios de los alumnos a la escuela mostrará muchos valores pequeños, en la mitad izquierda del histograma, son las de los alumnos que viven cerca y habrá pocos valores grandes de los alumnos que viven lejos.

17.2.3. Asimétrica a izquierda

Una **distribución es asimétrica a izquierda** cuando la mitad izquierda es más finita y más larga. En un **examen fácil,** la mayoría de las notas serán altas y estarán amontonadas del lado derecho, con unas pocas notas bajas (las del lado izquierdo).

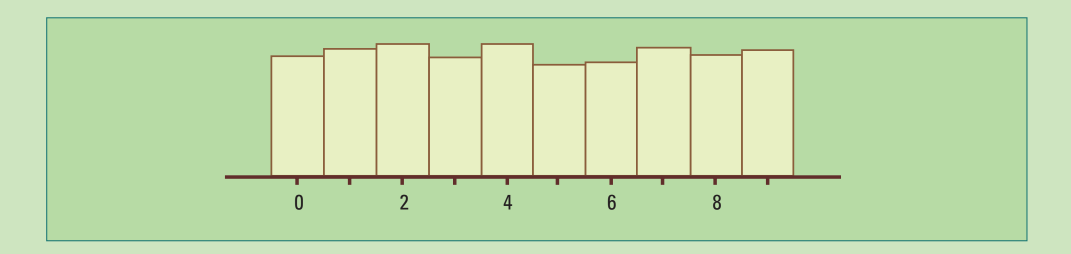


◻ 17.2.4. Con forma de campana

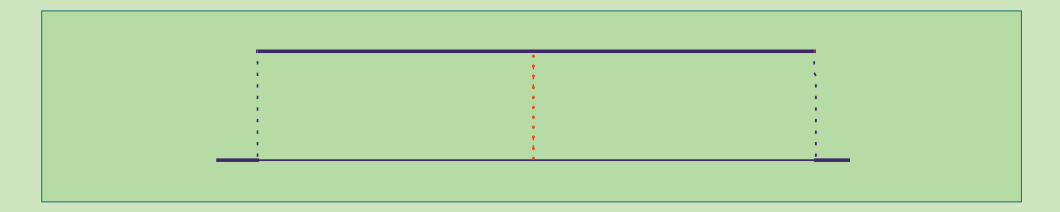


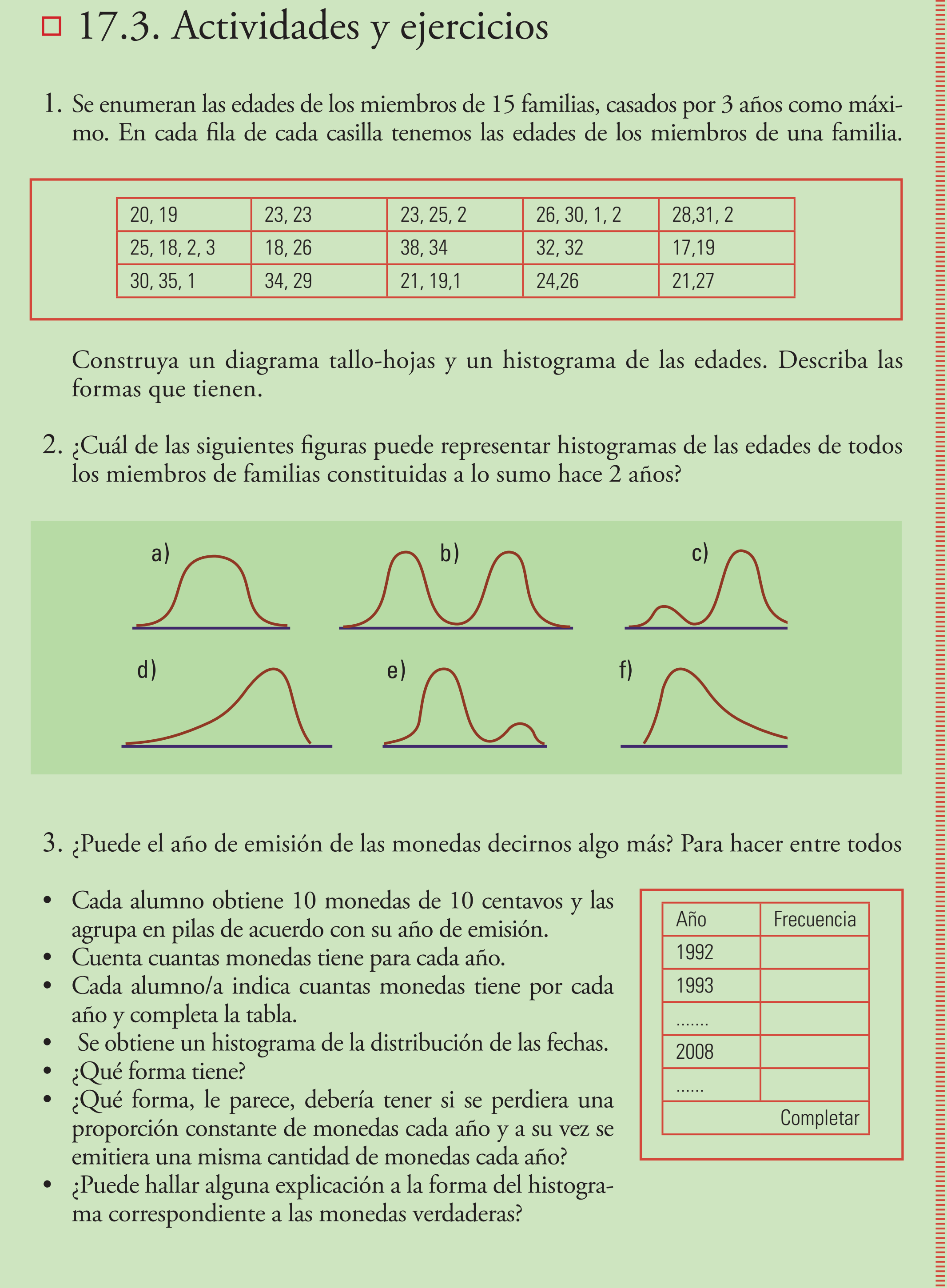
Una **distribución con forma de campana** es **simétrica** con un montículo en el centro y dos caídas como toboganes hacia los costados. Es una de las distribuciones de datos que tal vez aparezca con más frecuencia y es la más estudiada.

◻ 17.2.5. Uniforme



Las frecuencias de la última cifra de los resultados de una lotería muestran una distribución pareja sobre todos los dígitos de 0 a 9. Si el mecanismo que genera los números de la lotería funciona correctamente, ninguno de los dígitos tiene más chances de aparecer. Este tipo de distribuciones se llama uniforme y se representa mediante una recta:





17.3. Actividades y ejercicios

1. Se enumeran las edades de los miembros de 15 familias, casados por 3 años como máximo. En cada fila de cada casilla tenemos las edades de los miembros de una familia.

20, 19	23, 23	23, 25, 2	26, 30, 1, 2	28,31, 2
25, 18, 2, 3	18, 26	38, 34	32, 32	17,19
30, 35, 1	34, 29	21, 19,1	24,26	21,27

Construya un diagrama tallo-hojas y un histograma de las edades. Describa las formas que tienen.

2. ¿Cuál de las siguientes figuras puede representar histogramas de las edades de todos los miembros de familias constituidas a lo sumo hace 2 años?

3. ¿Puede el año de emisión de las monedas decirnos algo más? Para hacer entre todos

- Cada alumno obtiene 10 monedas de 10 centavos y las agrupa en pilas de acuerdo con su año de emisión.
- Cuenta cuantas monedas tiene para cada año.
- Cada alumno/a indica cuantas monedas tiene por cada año y completa la tabla.
- Se obtiene un histograma de la distribución de las fechas.
- ¿Qué forma tiene?
- ¿Qué forma, le parece, debería tener si se perdiera una proporción constante de monedas cada año y a su vez se emitiera una misma cantidad de monedas cada año?
- ¿Puede hallar alguna explicación a la forma del histograma correspondiente a las monedas verdaderas?

Año	Frecuencia
1992	
1993	
....	
2008	
......	
	Completar

4. Para estudiar las longitudes de las palabras, seleccione un artículo de una revista de deportes y otro de una de divulgación científica. Para cada uno de los artículos obtenga:

- la distribución de frecuencias
- la distribución de frecuencias relativas
- el histograma

de la variable "cantidad de letras" que tiene cada palabra. Compare las distribuciones obtenidas.

Observación: Diferentes idiomas tienen diferentes distribuciones de las longitudes de las palabras.

18. Medidas resumen

Media, mediana, rango, desvío estándar, distancia intercuartil

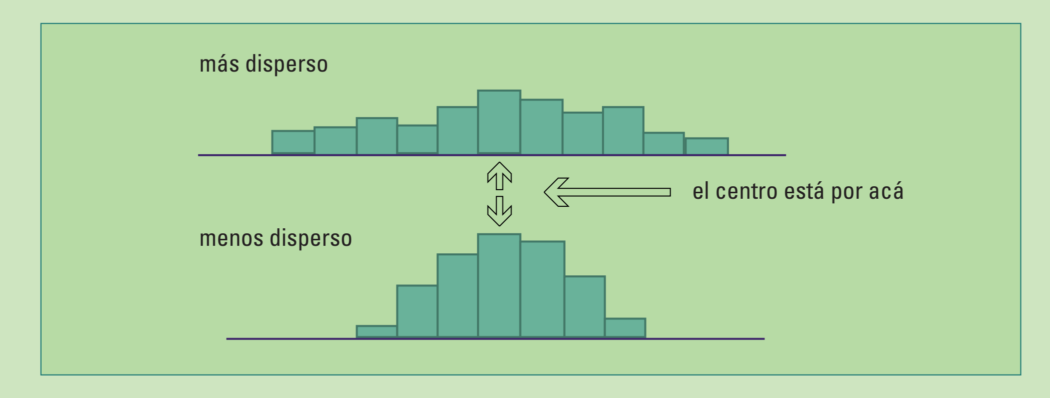
La mente humana puede captar la información que aportan diez números, cien es difícil y con mil, estamos perdidos. Por esa razón, es muy importante contar con pocos valores (medidas resumen), que de alguna manera deben describir las características más sobresalientes del conjunto que se está analizando.

Una medida resumen es un número. Se obtiene a partir de una muestra y, en cierta forma, la caracteriza. Es el valor de un estadístico. Por ejemplo, un porcentaje o una proporción son medidas resumen. Se utilizan con datos categóricos o con datos numéricos categorizados previamente. **Las medidas resumen permiten tener una idea rápida de cómo son los datos.** Pero, un estadístico mal utilizado puede dar una idea equivocada respecto de las características generales que interesa mostrar.

El cálculo de medidas resumen es el primer paso; se realiza cuando se recolectan los datos en un estudio para tener una idea de qué está pasando. Posteriormente, los investigadores pondrán a prueba sus hipótesis respecto a algún parámetro poblacional, estimarán características de la población y estudiarán posibles relaciones entre las variables. Cuando presentan sus conclusiones al público en general, las medidas resumen muestran los resultados en forma concisa y clara, volviendo a tener importancia.

En principio, se pueden obtener muchísimas formas de resumir los valores de un conjunto de datos numéricos. Es importante que sean fáciles de interpretar.

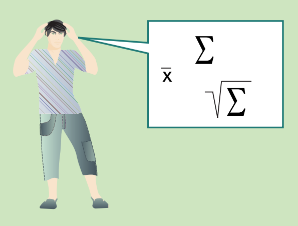
Cualquier conjunto de datos tiene **dos propiedades importantes:** un **valor central** y la **dispersión** alrededor de ese valor. Vemos esta idea en los siguientes histogramas hipotéticos:



Describiremos en este capítulo medidas de la posición del centro, la dispersión y otras medidas de posición. Veremos:

- Cómo se utilizan, en forma correcta o errónea.
- Qué significan.
- Qué dicen y qué no dicen estas medidas resumen.
- Cómo dependen de la distribución general de los datos.

Pero, a partir de ahora, además de gráficos necesitamos fórmulas.



Supongamos que tenemos un conjunto con n observaciones (datos), los representamos así:

$$x_1, x_2, x_3, \ldots, x_n$$

Se leen equis uno, equis dos, ..., equis ene y se pueden representar en una tabla:

(Número de) Observación	1 2 3 .... n
Valor	x_1 x_2 x_3 .... x_n

Ejemplo 18.1: Le preguntamos a 5 personas (n =5) cuántas cuadras camina por día y obtenemos.

Observación	1	2	3	4	5
Valor	4	15	8	31	17

Luego $x_1= 4$, $x_2=15$, $x_3= 8$, $x_4=31$, $x_5=17$

¿Cuál es el centro de estos datos? Respondemos esta pregunta en la siguiente sección.

☐ 18.1. Posición del centro de los datos

El **promedio** define el valor característico o central de un conjunto de números. Existen varios métodos para calcular el promedio. El método utilizado puede influir en las conclusiones. Cuando vemos un anuncio con la palabra promedio, debemos alertarnos porque quien lo ha escrito, probablemente eligió el método de cálculo para producir el resultado que le interesa marcar.

Veremos con detalle las dos formas principales para obtener un valor central o promedio:

- **La media:** Se obtiene sumando todos los valores del conjunto de datos y dividiendo la suma por la cantidad de datos en ese conjunto.
- **La mediana:** Es el valor central del conjunto de datos ordenados.

18.1.1. La media

La media se representa por $\bar{x}$ (equis raya o equis barra). Se obtiene sumando todos los datos y dividiendo por la cantidad total n de observaciones,

$$\bar{x} = \frac{\text{SUMA DE LOS DATOS}}{n}$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

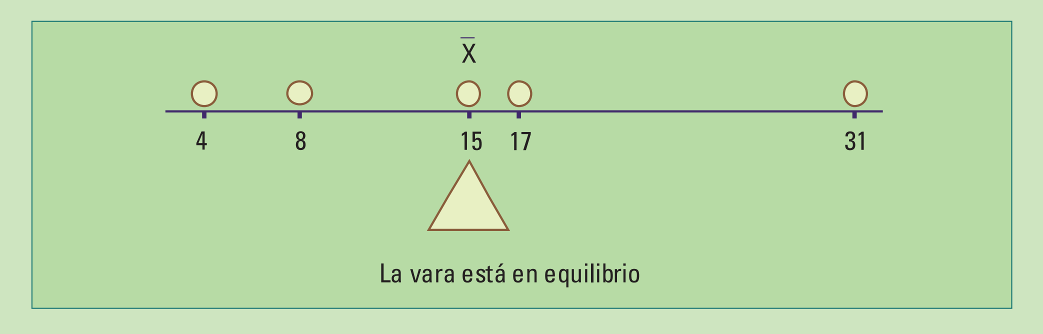
En el ejemplo anterior, la media de las cuadras caminadas por día es 15:

$$\bar{x} = \frac{4 + 15 + 8 + 31 + 17}{5}$$

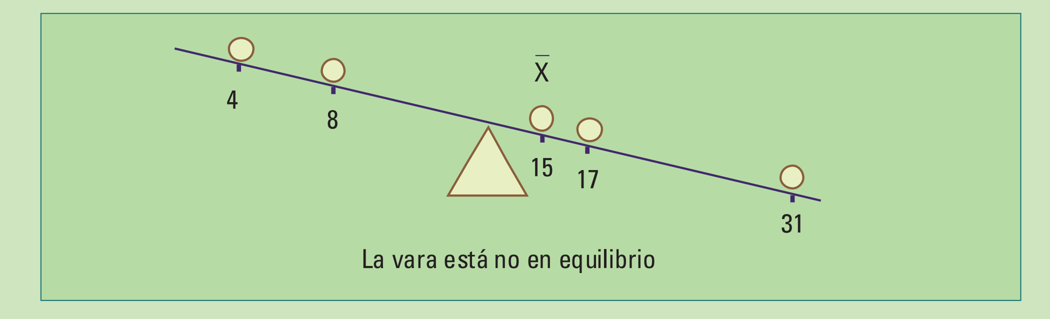
$$\bar{x} = \frac{75}{5}$$

$$\bar{x} = 15 \text{ CUADRAS}$$

Si sobre una vara numerada sin peso, se colocan pesos idénticos sobre el valor de cada dato, la **vara queda en equilibrio** cuando se la apoya en el punto correspondiente a la media.



La vara no queda en equilibro si se la apoya en cualquier otro punto.



Existe una abreviatura para la suma $x_1+x_2+...+x_n$. Se trata de la letra griega **sigma mayúscula** (comúnmente llamada **sumatoria**): $\sum$

En vez de la suma $x_1+x_2+...+x_n$ escribimos $\sum_{i=1}^{n} x_i$

y lo leemos como: "la suma de equis i, con i variando desde 1 hasta n".

Repito diez veces



$\sum_{i=1}^{n} x_i$ "La suma de x_i, con **i** variando desde 1 hasta **n**"

Así, la media de un conjunto de datos x_i es:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{ó} \quad \sum_{i=1}^{n} \frac{x_i}{n}$$

En el ejemplo 16.5, para los pesos de los 101 alumnos de 3 divisiones de 4to. Año, el peso medio es 58,90 kg:

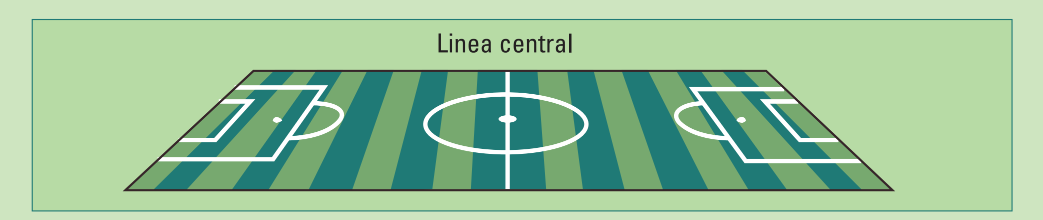
$$\sum_{i=1}^{101} \frac{x_i}{101} = \frac{5949}{101}$$

$$= 58,90 \text{ kg}$$



18.1.2. La mediana

La **mediana** es otro tipo de centro. Es el punto central de los datos, como la línea central que divide el campo de juego de fútbol en dos partes iguales.



La mediana deja la misma cantidad de datos a cada lado.

Para hallar la mediana del conjunto de datos (4, 15, 8, 31, 17) del ejemplo 18.1:

- Primero los ordenamos de menor a mayor (4, 8, 15, 17, 31).
- Luego, la mediana es el valor central (15).

Para las cuadras que caminan por día las cinco personas elegidas al azar, el valor central, la mediana, es15. Quedan dos datos a cada lado de la mediana. En este ejemplo, la media coincide con la mediana, pero puede no ocurrir.

4 8 (15) 17 31

Si la cantidad de datos es **par** (4, 15, 8, 17) no hay una observación central, sino **un par** de **observaciones centrales** (8 y 15). La mediana (11,6) es el promedio de estos dos valores.

4 8 15 17 promediamos el 8 y el 15 $\frac{8+15}{2} = 11{,}6$

La regla general para calcular la mediana de n datos ordenados es:

- Si la **cantidad de datos** es **impar,** la mediana es el valor del centro, se encuentra en la posición (n+1)/2.
- Si la **cantidad de datos** es **par,** la mediana es el promedio de los dos valores centrales, se encuentran en las posiciones n/2 y (n/2)+1

Para los datos de los pesos de los 101 alumnos (ejemplo 16.5) la mediana es 58 kg. Como ya hemos construido el diagrama tallo hoja ordenado, la obtenemos directamente contando desde el dato más pequeño hasta el dato en la posición 51 (51=101+1)/2):

3 |
3 | 78
4 | 2334
4 | 566788888
5 | 0000001111122222222223444444
5 | 5556677788**8**99 ← Aquí se encuentra la mediana
6 | 011223333444
6 | 555666667777778889999
7 | 00112234
7 | 99
8 | 1
8 | 5

Pruebe contar 51 desde el dato más grande hacia los más chicos; la mediana también da 58.

18.1.3 ¿Por qué utilizamos más de una medida de posición del centro de los datos?

Cada una de las dos medidas presentadas tiene ventajas y desventajas.

La media utiliza todos los datos para su cálculo. Si los datos presentan un histograma simétrico calcular **la media es lo mejor** para obtener el centro de los datos, en este caso la mediana será muy parecida.

Siguiendo con el ejemplo 18.1 (cuadras que caminan por día 5 personas) la media y la mediana coinciden.

4   8   (15)   17   31   mediana = media

La mediana no se verá afectada si los datos presentan algún **valor atípico (316),** es decir, un dato alejado del patrón general (también llamado **outlier** en inglés), mientras que la media sí.

4   8   15   17   316 ⇐ valor atípico

El outlier puede ocurrir si una de las personas entrevistadas tiene hábitos diferentes a lo habitual (316 en lugar de 31), o si cometimos un error. La mediana seguirá siendo 15, pero la media será 72. ¿Es razonable decir que 72 cuadras por día en promedio representan las distancias caminadas por la mayoría de las personas?

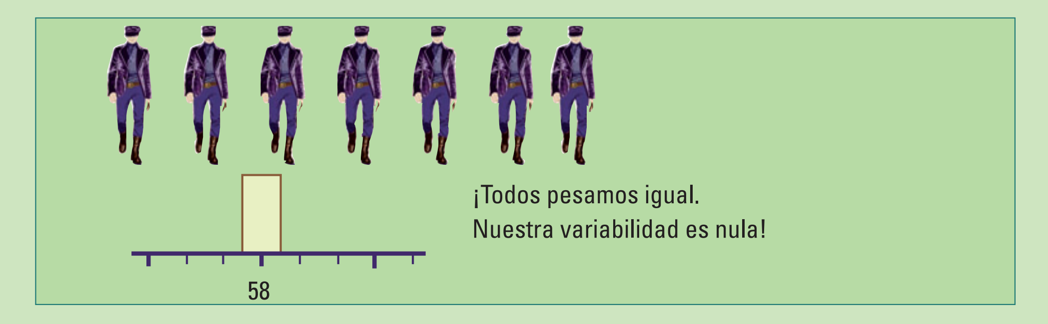
4   8   15   17   316

mediana ⇗   media = 72 ⇗

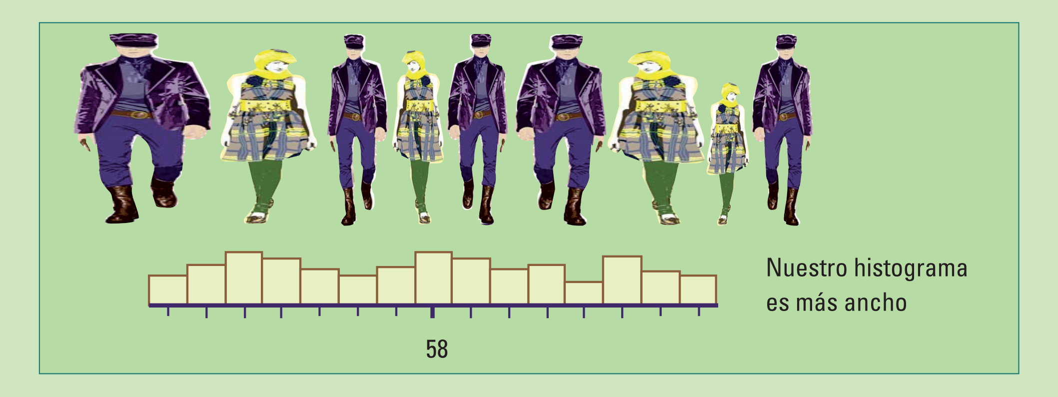
La media (72) ya no representa a la mayoría de los datos, por eso, decimos que **la media es sensible ante la presencia de valores atípicos (outliers).**

□ 18.2. Medidas de dispersión o variabilidad

Si todos los alumnos pesaran 58 kg, tendríamos un conjunto de datos iguales.



Otro conjunto de alumnos con mediana igual a 58kg podría tener pesos diferentes y los datos estarían más dispersos.



Además de conocer el punto central de un conjunto de datos, también nos interesa describir su dispersión, es decir cuán lejos tienden a estar los datos de su centro.

La variabilidad está presente en todos los conjuntos de datos. Sea cual fuere la característica, es casi imposible que dos mediciones sean idénticas. Esto se debe a que:

- Diferentes individuos tienen diferentes características (peso, altura, inteligencia, glóbulos rojos en sangre), al cuantificarlas resultan en valores diferentes de las variables correspondientes.
- Diferentes mediciones de una misma característica dan como resultado diferentes valores debido al inevitable error de medición.

Los métodos estadísticos son imprescindibles para analizar los datos debido a su variabilidad. El truco consiste en tener medidas que la capten de la mejor manera posible.

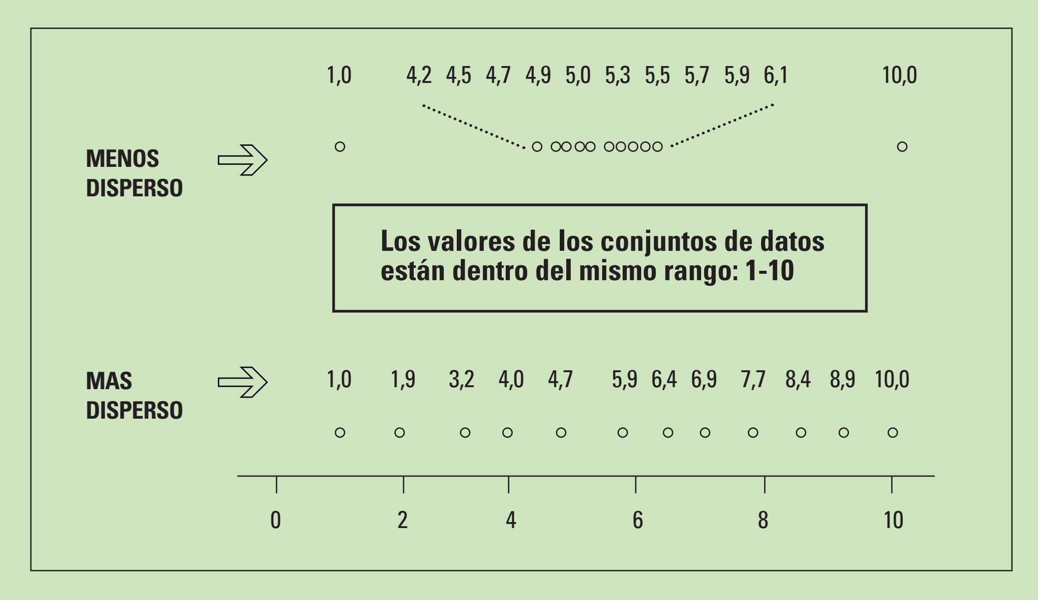
18.2.1. Rangos y distancia intercuartil

El rango de valores donde se encuentran los datos permite apreciar su variabilidad o dispersión (cuán desparramados están).

La medida natural para evaluar dicha dispersión es la distancia entre el valor mínimo y el valor máximo de los datos (máximo-mínimo).

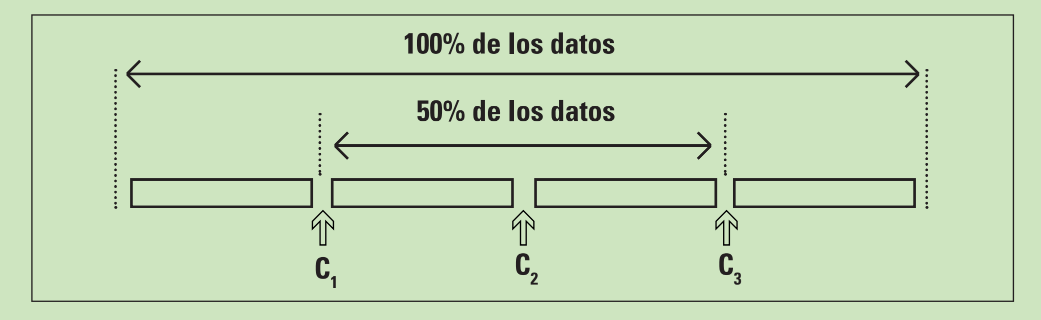
Tiene algunos inconvenientes:

- Es muy sensible a la presencia de valores atípicos.
- Como utiliza sólo dos datos, no puede distinguir dos conjuntos con máximos y mínimos coincidentes, pero uno tendrá la mayoría de sus valores mucho más concentrados que el otro.



La figura representa a los siguientes conjuntos de datos {1,0 4,2 4,5 4,7 4,9 5,0 5,3 5,5 5,7 5,9 6,110,0 } y {1,0 2,9 3,5 4,0 4,7 5,9 6,4 6,9 7,7 8,4 8,9 10,0}. La mayoría de los valores del primer conjunto están más concentrados que la mayoría del segundo conjunto pero tienen el mismo rango. El rango en este caso no distingue dos conjuntos de datos con diferentes dispersiones.

Para corregir los problemas se utiliza la distancia entre el valor mínimo y el valor máximo del 50% central de los datos, llamada distancia intercuartil.



¿Cómo se calcula la distancia intercuartil?:

1. Se ordenan los datos.
2. Se calcula la mediana (C_2), que los divide en 2 partes con igual cantidad de datos de cada lado.
3. Se calcula la mediana de la mitad más baja (grupo inferior), es el cuartil inferior (C_1)
4. Se calcula la mediana de la mitad más alta (grupo superior), es el cuartil superior (C_3)
5. La distancia intercuartil (DIC) es la diferencia entre el cuartil superior y el cuartil inferior: **$DIC = C_3 - C_1$**

Cuando la mediana coincide con uno de los datos se la puede considerar parte de los dos grupos, el superior y el inferior (esta regla es arbitraria y algunos autores no la cuentan en ninguno de los dos).

¿Qué mide la distancia intercuartil?
Como medida de dispersión, la distancia intercuartil mide la longitud del intervalo en el cual se encuentra el 50% central de los datos. Cuanto más dispersos estén los datos, mayor será la distancia intercuartil.

Nuevamente, consideremos los pesos de los 101 alumnos (ejemplo 16.5). La mediana está en la posición 51 y vale 58 kg. Para hallar el cuartil inferior calculamos la mediana de los 51 valores más chicos. Se encuentra en la posición (51+1)/2=26. Contamos 26 lugares desde los más chicos y obtenemos el valor 51 kg del cuartil inferior.

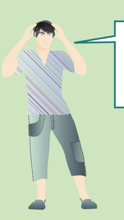


No confundir la posición 51 (donde se encuentra la mediana) **con 51 kg,** el valor del cuartil inferior que se encuentra en la posición 26.

Contando 26 lugares desde los **valores más altos** obtenemos el valor 67 kg del cuartil superior.

La distancia intercuartil se obtiene como la diferencia entre el cuartil superior y el cuartil inferior (DIC = 67 kg - 51 kg = 16), es la diferencia entre la mediana de los alumnos más pesados y la mediana de los más livianos. El 50% de los pesos difieren a lo sumo en 16 kg. El 50% de los pesos están entre 51 kg y 67 kg.

3 |
3 | 78
4 | 2334 **Cuartil inferior**
4 | 566788888 ↓
5 | 0000001111**1**222222223444444
5 | 5556677788**8**99 ← **Aquí se encuentra la mediana**
6 | 011223333444
6 | 555666667**7**7777788899999
7 | 00112234 ↑
7 | 99 **Cuartil superior**
8 | 1
8 | 5



La mediana esta en la posición 51 y tiene un valor de 58 kg.
El cuatil inferior se encuentra en la posición 26 y tiene un valor de 51 kg.

No confundir la posición de un dato con el valor de un dato.

18.2.2. Los cinco números resumen y el gráfico de caja y brazos

El mínimo, el cuartil inferior, la mediana, el cuartil superior y el máximo son cinco números. Dan una idea de cómo está distribuido un conjunto de datos. Se los llama los cinco números resumen y se los representa por:

Mínimo C_1 M C_3 Máximo

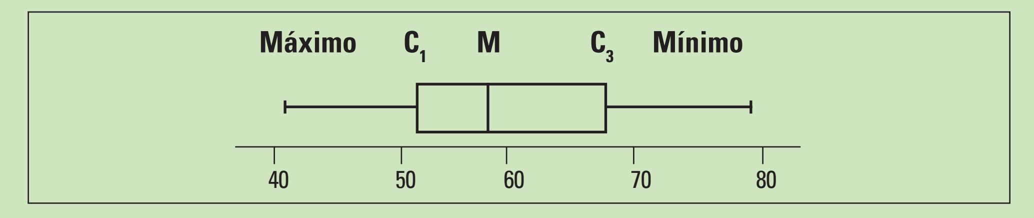
El 50% de los datos se encuentran entre el cuartil inferior y el superior.

Los cinco números resumen de los pesos de los alumnos de 4to. año son:

Mínimo	C_1	M	C_3	Máximo
37	51	58	67	85

El 50% de los alumnos tiene un peso entre 51 y 67 kg.

Los cinco números resumen se representan gráficamente en un Gráfico de caja (Box-plot).

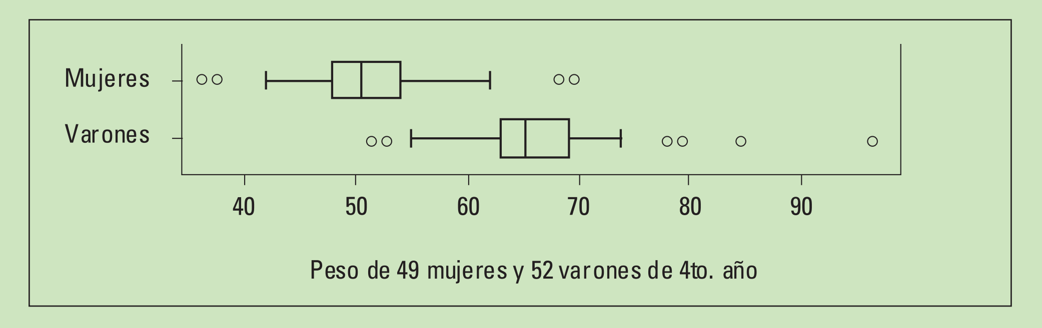


Los cuartiles forman los bordes de la caja y la mediana está dentro de la caja. Dos líneas - los brazos- se extienden, una desde cada borde de la caja, hasta el dato con valor máximo y mínimo respectivamente, mientras no sean valores atípicos (es decir, se encuentren dentro de 1,5 DIC).

Si agregamos un peso de 97 kg a los datos de los pesos, el boxplot muestra un valor atípico.



Los gráficos caja sirven especialmente cuando queremos comparar varios conjuntos de datos. En el ejemplo de los pesos, comparemos los de varones y de mujeres.



Entre las mujeres hay 2 que pesan menos que la mayoría y otras 2 más (por fuera de los brazos). Entre los varones se detectan 2 en los valores menores y 4 en los valores mayores. El 75% de las mujeres son más livianas que los hombres (excluyendo los 2 valores atípicos bajos de los hombres). Los **cinco números resumen** muestran los detalles:

Peso de Mujeres

Mínimo	Cuartil inferior - C_1	Mediana	Cuartil superior - C_3	Máximo
37	48	51	54	70

Peso de Varones

Mínimo	Cuartil inferior - C_1	Mediana	Cuartil superior - C_3	Máximo
51	63	66	69	97

18.2.3. Desvío estándar

La descripción de una distribución mediante medidas resumen es utilizada desde hace muchísimos años. Pero, la propuesta de utilizar los 5 números resumen es relativamente nueva. Fue hecha por John Tukey por los años 70, cuando comenzaban a utilizarse las computadoras.

La mediana y los cuartiles son muy sencillos de calcular a mano cuando la cantidad de datos es relativamente pequeña. Cuando se tienen muchos datos, la dificultad se encuentra en ordenarlos. Por esa razón, aunque la mediana era conocida casi no se utilizaba antes del advenimiento de las computadoras.

La media, es mucho más **fácil de calcular a mano** cuando hay muchos datos. Sólo requiere del uso de operaciones aritméticas, para hallar un número representativo de la mayoría de los datos.

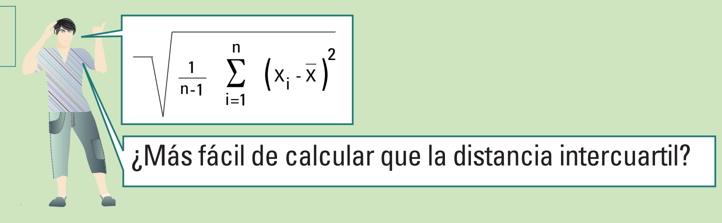
El **desvío estándar** es una **medida** de dispersión **basada en la media** y **utiliza todos los datos.** Durante muchos años la **media y el desvío estándar** fueron, y tal vez sigan siendo, las **medidas resumen más utilizadas.**

El desvío estándar representa una distancia típica de cualquier punto del conjunto de datos a su centro (medido por la media). Es una distancia promedio de cada observación a la media.

El desvío estándar de los datos de toda una población (desvío estándar poblacional) se denota con la letra griega σ (sigma minúscula). Pero la mayoría de las veces los parámetros poblacionales son desconocidos. ¿Qué se hace? Se calcula un estimador (s, desvío estándar muestral) utilizando una muestra.

> La distinción entre el desvío estándar poblacional y el desvío estándar muestral vale para todos los estadísticos descriptos (media, mediana, cuartiles, distancia intercuartil, etc.). Tal como vimos en los capítulos 9 y 10, si el cálculo de un estadístico se realiza utilizando una muestra para estimar un parámetro, el resultado tendrá un error de muestreo.

¡Desvío estándar!



El desvío estándar se calcula promediando la diferencia entre cada dato y la media, elevadas al cuadrado. Como este resultado tiene las unidades al cuadrado, luego se saca la raíz cuadrada.

Para un conjunto de n datos:

1. Se calcula la distancia de cada dato a la media: $x_i - \bar{x}$
2. Se eleva al cuadrado: $(x_i - \bar{x})^2$
3. Se promedie dividiendo por n-1 y, así, se obtiene la varianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

4. Por último se calcula la raíz cuadrada

$$s = \sqrt{s^2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Para el conjunto de datos de las cuadras que 5 personas caminan por día, del ejemplo 18.1 (x_1=4, x_2=15, x_3=8, x_4=31, x_5=17, n=5 y $\bar{x}$=5), la varianza muestral es 107,5 cuadras[2]:

$$s^2 = \frac{(4-15)^2 + (15-15)^2 + (8-15)^2 + (31-15)^2 + (17-15)^2}{(5-1)}$$

$$s^2 = \frac{121 + 0 + 49 + 256 + 4}{4}$$

$$s^2 = 107{,}5$$

Cuanto más grande es la varianza muestral, más dispersos están los datos. Una medida de dispersión debe tener las mismas unidades que los datos.

La varianza muestral, en nuestro ejemplo está en cuadras al cuadrado, entonces por supuesto, debemos sacar la raíz cuadrada.

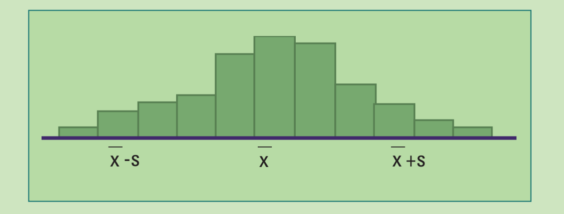
El desvío estándar es 10,37 cuadras:

$$s = \sqrt{107{,}5}$$

$$s = 10{,}37$$

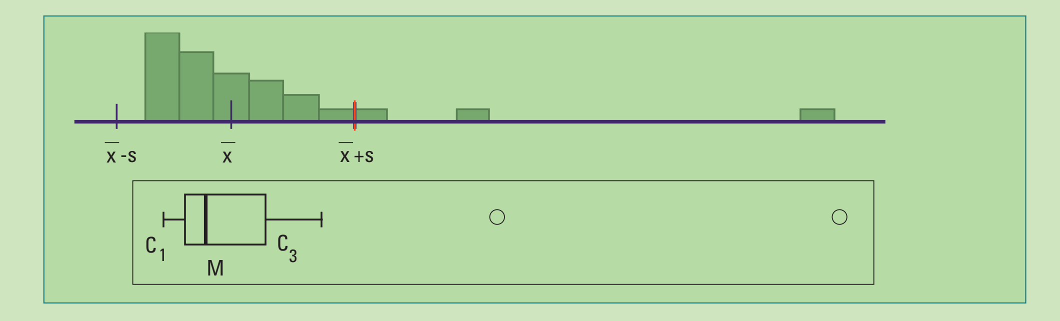
18.3. Centro y dispersión en diferentes tipos de distribuciones

La media y el desvío estándar son muy buenos para resumir datos con histogramas razonablemente simétricos y sin valores atípicos.



Sin embargo, la media y el desvío estándar no son una buena representación de la distribución de datos cuando tienen **valores atípicos** o sus **histogramas son asimétricos.**

La figura siguiente muestra el histograma de un conjunto de datos con distribución **asimétrica a derecha.** En este caso, **es mejor utilizar** la mediana y la distancia intercuartil, y mejor aún, **los 5 números resumen.**



El intervalo con extremos en **x̄-s** y **x̄+s no es una buena representación de los datos: x̄-s** se encuentra fuera del rango de los valores observados (está a la izquierda del valor más pequeño) y quedan valores a la derecha de **x̄+s**. El gráfico caja (boxplot) describe más precisamente el rango donde se encuentran los datos. El rango intercuartil que forma la caja contiene el 50% de los datos y los brazos se extienden hasta el último dato de cada lado. Se distinguen dos datos atípicos (en inglés: outliers, significa: yacen fuera).

En el ejemplo siguiente mostramos cómo las medidas resumen pueden contar una parte muy parcial de la historia.

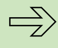
Ejemplo. "Admítelo una salchicha no es una zanahoria". Así decía la revista "El Consumidor" en un comentario sobre la baja calidad nutricional de las salchichas.
(Introduction to the practice of Statistics Moore mc Cabe pág. 28).

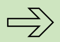
Hay tres tipos de salchichas:
1. carne vacuna,
2. mezcla (carne porcina, vacuna y de pollo)
3. pollo.

¿Existe alguna diferencia sistemática entre estos tres tipos de salchichas, en estas dos variables? Mirar directamente los datos sirve de muy poco.

Calorías y sodio en salchichas por tipo. Tabla 18.1

Vacuno		Mezcla		Pollo	
Calorías	**Sodio**	**Calorías**	**Sodio**	**Calorías**	**Sodio**
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	528
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	344	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522





Vacuno		Mezcla		Pollo	
Calorías	Sodio	Calorías	Sodio	Calorías	Sodio
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	296				
132	253				

Comparemos la cantidad de calorías entre los tres tipos de salchichas utilizando gráficos caja. Recordemos que están basados en los números resumen:

	Mínimo	Cuartil inferior C_1	Mediana	Cuartil superior C_3	Máximo
Vacuno	111	140,5	152,5	178,5	190
Mezcla	107	139	153	179	195
Pollo	86	102	129	143	170

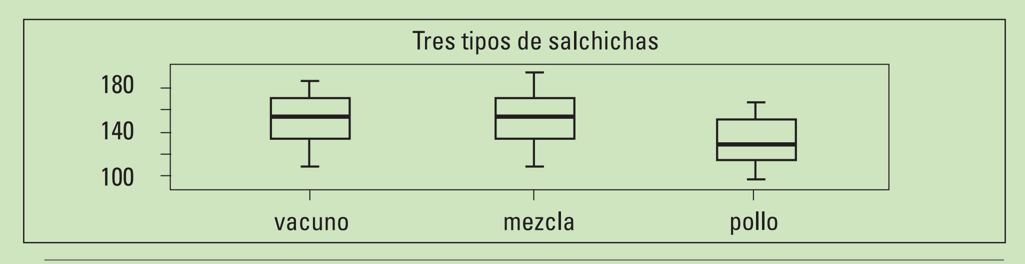


Figura 18.1. *Gráficos caja de la cantidad de sodio de tres tipos de salchichas.*

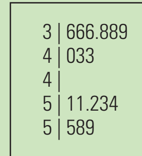
Vacuno	Mezcla	Pollo
8 |	8 |	8 | 67
9 |	9 |	9 | 49
10 |	10 | 7	10 | 226
11 | 1	11 |	11 | 3
12 |	12 |	12 | 9
13 | 1259	13 | 5689	13 | 25
14 | 1899	14 | 067	14 | 2346
15 | 2378	15 | 3	15 | 2
16 |	16 |	16 |
17 | 56	17 | 2359	17 | 0
18 | 146	18 |	18 |
19 | 00	19 |	19 |

Figura18.2. *Diagramas tallo hoja de la cantidad de sodio de tres tipos de salchichas. La coma decimal se encuentra un dígito a la derecha de la barra vertical* (|).

Vemos una tendencia general en las salchichas de pollo a presentar menor cantidad de calorías. Pero nos perdemos los detalles.

Los diagramas tallo hoja de las salchichas de carne vacuna y mezcla (figura 18.2) muestran la presencia de 2 grupos, y un valor aislado en la cola inferior. Sin embargo, como cada cuartil se encuentra aproximadamente en el centro de cada uno de los dos grupos, la distancia intercuartil refleja la distancia entre los grupos y, por lo tanto, el valor inferior no es detectado como dato atípico.

Analicemos ahora la distribución de la **cantidad de sodio** en las salchichas de **pollo** (tabla 18.1), cuyo diagrama tallo hoja tenemos a continuación


3 | 666.889
4 | 033
4 |
5 | 11.234
5 | 589

Tanto el diagrama tallo hoja como el histograma (figura 18.3) revelan la presencia de dos grupos:

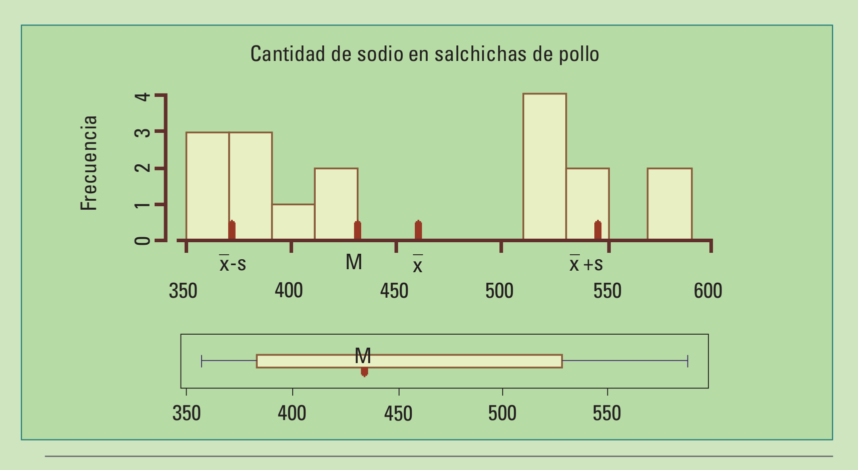


***Figura18.3.** Histograma y gráfico caja de la cantidad de sodio de salchichas de pollo de diferentes marcas.*

Los valores ordenados de la cantidad de sodio en salchichas de pollo son:
357 358 359 375 383 387 396 426 430   513 522 528 542 581 588

La media (449,66) se encuentra fuera de los datos, la mediana (426) cerca del borde de uno de los dos grupos. El intervalo ($\bar{x}$-s, $\bar{x}$+s) no es una buena representación de los datos y el gráfico caja tampoco.

Recomendamos realizar gráficos caja fundamentalmente para comparar la distribución de varios conjuntos de datos. Un diagrama de tallo y hojas o un histograma son mejores para analizar la distribución de datos de una única variable. Generalmente, los detalles agregan poco, pero es importante estar preparados para las ocasiones en que sí agregan mucho.

El significado de las medidas resumen está atado a la forma de la distribución de los datos. Esto tiene especial importancia con el desvío estándar pues se utiliza muchísimo en las descripciones de los datos. Su fama se debe a la íntima conexión que tiene el desvío estándar con la curva de Gauss. Lo veremos en el capítulo 20.

El **desvío estándar no significa nada** si los datos **no son Normales** ni aproximadamente Normales.

La **media no describe el centro** si los datos **no son simétricos**.

La mediana y la distancia intercuartil pueden fallar si los datos forman grupos.

18.4. Actividades y ejercicios

En los ejercicios 1-4 indique cual es la respuesta correcta o la afirmación que completa la frase. Explique brevemente

1. ¿Cuál de las siguientes opciones da la mejor descripción de los datos cuando estos presentan intervalos vacíos y grupos?

a) La media y el desvío estándar.
b) La mediana y el rango intercuartil.
c) Un gráfico caja con los 5 números resumen.
d) La mediana y el rango.
e) Un diagrama tallo-hoja o un histograma.
f) Ninguno de los anteriores permite mostrar intervalos vacíos y grupos.

2. ¿Cual de las siguientes medidas de posición y variabilidad son adecuadas cuando se sospecha la presencia de datos atípicos?

a) La media y el desvío estándar.
b) La media y el máximo menos el mínimo.
c) La media y la distancia intercuartil.
d) La mediana y la distancia intercuartil.
e) La mediana y el máximo menos el mínimo.

3. Si el desvío estándar de un conjunto de datos es cero, se puede concluir que:

a) La media es cero.
b) La mediana es cero.
c) Todos los datos valen cero.
d) Hay un error de cálculo.
e) La media mayor que la mediana.
f) Todos los datos son iguales.

4. Si el 20% de los datos se encuentra entre 10 y 40. Si se dividen por dos todos los valores y luego se les suma 10, también a todos, entonces:

a) El 10% de los datos resultantes estarán entre 15 y 30.
b) El 20% de los datos resultantes estarán entre 15 y 30.
c) El 15% de los datos resultantes estarán entre 15 y 30.
d) El 10% de los datos resultantes estarán entre 5 y 20.
e) El 15% de los datos resultantes estarán entre 5 y 20.
f) El 20% de los datos resultantes estarán entre 5 y 20.

5. Lleve una **balanza** a su división y **registre el peso y la edad** de todos los alumnos y alumnas.

   a. Describa, utilizando histogramas, cómo se distribuyen los **pesos** de todos, juntos y separados, varones y mujeres. Utilice también medidas resumen: media o mediana; distancia intercuartil desvío estándar. Indique cuales son las más adecuadas.
   b. Describa como se distribuyen las **edades** de todos juntos y separados, varones y mujeres. Utilice herramientas gráficas para comparar y también medidas resumen: media o mediana; distancia intercuartil o desvío estándar. Indique cuáles son las más adecuadas.

6. Lleve la **balanza** a **2 divisiones** de **años anteriores** y registre el peso y la edad de todos los alumnos y alumnas.

   a. Describa como se distribuyen los pesos de todos juntos y separados, varones y mujeres. Utilice herramientas gráficas para comparar y también medidas resumen: media o mediana; distancia intercuartil o desvío estándar. Indique cuáles son las más adecuadas.
   b. Describa como se distribuyen las **edades** de todos juntos, varones y mujeres. Utilice herramientas gráficas para comparar y también medidas resumen: media o mediana; distancia intercuartil o desvío estándar. Indique cuáles son las más adecuadas.

Compare los resultados de los distintos años.

7. Realice una encuesta en **su división** para averiguar la cantidad de horas que dedica cada alumno a estudiar y a mirar televisión.

   a. Pregúntele a todos y tendrá datos poblacionales para su división.
   b. ¿Le parece que esa muestra es representativa de todos los alumnos de la escuela?
   c. Elija las variables más relevantes para su encuesta. Establezca las preguntas y evalúe si estas pueden producir sesgo en las respuestas.
   d. Compare cómo se distribuyen las horas entre varones y también entre mujeres.
   e. Utilice herramientas gráficas para comparar y también medidas resumen. Media o mediana. Distancia intercuartil o desvío estándar. Indique cuales son las más adecuadas.

8. Realice una encuesta **en toda su escuela** para averiguar la cantidad de horas que dedica cada alumno a estudiar y a mirar televisión. Utilice **una muestra representativa de todos los años y de género,** especifique como la elegirá. Para esta encuesta puede utilizar las mismas variables y las preguntas que utilizó para su división o modificarlas, según se consideren adecuadas a la luz de los resultados obtenidos. Puede utilizar la colaboración de algún alumno de cada división.

19. Otras medidas de posición: los percentiles

El cuartil inferior es el percentil 25.
La mediana es el percentil 50.

Los percentiles nos permiten responder preguntas como:

- Tiene 16 años, mide 1,57 cm, es la primera de la fila de su división, ¿significa que es petisa?
- Los alumnos de 2do. año de una escuela practican deportes 6 horas por semana ¿es mucho o poco en comparación con otras escuelas?
- Tiene 25 años, se tomó la presión y obtuvo 130/70 (sistólica /diastólica). ¿Es normal?
- ¿Cuál es la longitud debajo de la cual se encuentran el 90% de los bebés recién nacidos?
- Un bebé nació con 3800 g y 50 cm de longitud corporal ¿No está un poco gordito? ¿Será petiso?
- En relación al volumen de ventas, ¿cómo está posicionada una compañía azucarera en comparación con la competencia?

Los percentiles se utilizan para determinar la posición relativa, en porcentaje, de la posición que ocupa un valor dado, de una variable, en relación a todos los valores de la misma en un grupo o en una población.



PERCENTILES ¿%? ¿Valores? ¿Valor fijo?

Ejemplo: La tabla siguiente muestra los percentiles de la altura (m) de mujeres y varones de 16 años (Gráfico Nº 6 Guías para la Evaluación del Crecimiento. Sociedad Argentina de Pediatría. 2001).

Percentiles de la altura (m) de mujeres y varones de 16 años. TABLA 19.1

Percentil	**3**	10	**25**	50	75	90
Mujer	**1,49**	1,53	**1,56**	1,60	1,64	1,68
Varón	1,56	1,60	1,65	1,70	1,74	1,79

Fuente. *Guías para la Evaluación del Crecimiento. Sociedad Argentina de Pediatría. 2001.*

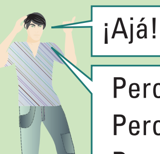
El 3% de las jóvenes de 16 años miden menos o igual que 1,49 m. El **percentil 3 de la altura** de las jóvenes de 16 años es de **1,49 m.**

El 25% de las jóvenes de 16 años miden menos o igual que 1,56 m. El **percentil 25 de la altura** de las jóvenes de 16 años es de **1,56 m.**

El 50% de las jóvenes de 16 años miden menos o igual que 1,60 m. El **percentil 50 de la altura** de las jóvenes de 16 años es de **1,60 m.**

En general hablaremos del **percentil K**. Si decimos percentil 25, eso significa K = 25. También se lo denomina **percentil del K** %, se dice "percentil del 25%" en lugar de "percentil 25".

- ¿A qué población corresponden los percentiles de la tabla 18.1? A mujeres y varones de 16 años de la Argentina.
- ¿Cuál es el percentil 10 de los varones de 16 años? El percentil 10 de la altura de los varones es 1,60 m ¿Qué significa ese valor? Significa que el 10 % de los jóvenes de 16 años miden menos o igual que 1,60 m.
- ¿Cuál es el percentil 10 de las mujeres de 16 años? El percentil 10 de la altura de los varones es 1,53 m ¿Qué significa ese valor? Significa que el 10 % de las jóvenes de 16 años miden menos o igual que 1,53 m.
- ¿Son iguales los valores anteriores? ¿Por qué? Los percentiles 10 de varones y mujeres difieren porque las distribuciones de las alturas son diferentes.



¡Ajá!

Percentil 25 ¡Es el cuartil inferior!
Percentil 50 ¡Es la mediana!
Percentil 75 ¡Es el cuartil superior! y hay muchos más!!!

En general: Una proporción p de jóvenes de 16 años tiene una altura por debajo del percentil 100 x p de la altura de las jóvenes de 16 años.

Más en general: Una proporción p de observaciones de una variable está por debajo del percentil 100 x p de dicha variable.

"Mido 1,57 m. ¡No soy petisa! El 25 % de las chicas de mi edad miden menos que yo."

19.1. ¿Cómo se calcula un percentil en un conjunto de datos?

19.1.1. Cuando los datos no están agrupados

Veamos ahora una **forma general para hallar el percentil K para cualquier conjunto de datos:**

Paso 1. Ordene los datos de menor a mayor
Paso 2. Calcule **K/100** y multiplíquelo por la cantidad total de datos **n.**
Paso 3. Redondee n K/100 al entero más cercano.
Paso 4. Cuente desde el dato más chico hacia el más grande tantos lugares como el número hallado en el paso 3.

Ejemplo: Ahora usaremos **peso,** no altura.

Retomemos nuevamente los datos del ejemplo 16.5. Consideremos los **pesos** (en kg) de las 49 alumnas de 4to. año y hallemos el percentil 40.

Paso 1. Ordenamos los datos:
37 38 42 43 43 44 45 46 46 47 48 48 48 48 48 50 50 50 50 50 50 51 51 51 51
52 52 52 52 52 52 52 52 54 54 54 54 54 54 55 55 56 56 57 58 60 62 68 70

Paso 2. Calculamos **20/100** = 0,2 y lo multiplicamos por la cantidad total de datos 49. Esto da como resultado 9,8

Paso 3. Redondeamos 9,8 al número entero más cercano, o sea 10.

Paso 4. Contamos desde el dato más chico hacia el más grande 10 lugares:

37 38 42 43 43 44 45 46 **46** 47 48 48 48 48 48 50 50 50 50 50 50 51 51 51 51
52 52 52 52 52 52 52 52 54 54 54 54 54 54 55 55 56 56 57 58 60 62 68 70

No confundir el valor del percentil con el porcentaje.

Por lo tanto, 46 kg es el percentil 20 para los datos de los pesos (en kg) de las 49 alumnas de 4to. año; 46 es el valor y 20 es el porcentaje.

19.1.2. Cuando los datos están agrupados

Cuando los datos tienen muchos **valores repetidos** es más conveniente utilizar una tabla de frecuencias para calcular los percentiles. Utilizaremos nuevamente las 49 alumnas de 4to. año. para calcular el percentil 20. Si consideramos **todos los valores desde el más chico hasta 46** (tabla 19.2) se acumulan **aproximadamente** el 20 % de los pesos.

Ejercicios utilizando la tabla 19.2:

- Halle el peso correspondiente al percentil 90. Si no lo encuentra exactamente, obtenga el más cercano.
- Una alumna pesa 52 kg, ¿en qué percentil se encuentra? ¿Es un percentil respecto de las 49 alumnas o respecto a toda la población?

Solución

- El porcentaje acumulado más cercano a 90 es 89,80 y le corresponde un peso de 57 kg. Podemos decir que aproximadamente el 90% de las alumnas del curso pesa a lo sumo 57 kg.
- Se encuentra en el percentil del 67,35 %. Se trata de un percentil respecto a las 49 alumnas del curso.

FRECUENCIAS DEL PESO DE LAS 49 ALUMNAS DE 4TO. AÑO. TABLA 19.2

Peso	Frecuencia	Frec. acumulada	Frec. relativa (en %)	Porcentaje acumulado
37	1	1	2,04	2,04
38	1	2	2,04	4,08
39	0	2	0,00	4,08
40	0	2	0,00	4,08
41	0	2	0,00	4,08
42	1	3	2,04	6,12
43	3	6	6,12	12,24
44	1	7	2,04	14,29
45	1	8	2,04	16,33
46	**2**	**10**	**4,08**	**20,41**
47	1	11	2,04	22,45
48	5	16	10,20	32,65
49	0	16	0,00	32,65
50	5	21	10,20	42,86
51	4	25	8,16	51,02
52	8	33	16,33	67,35
53	0	33	0,00	67,35
54	6	39	12,24	79,59
55	2	41	4,08	83,67
56	2	43	4,08	87,76
57	1	44	2,04	89,80
58	1	45	2,04	91,84
59	0	45	0,00	91,84
60	1	46	2,04	93,88
61	0	46	0,00	93,88
62	1	47	2,04	95,92
63	0	47	0,00	95,92
64	0	47	0,00	95,92
65	0	47	0,00	95,92
66	0	47	0,00	95,92
67	0	47	0,00	95,92
68	1	48	2,04	97,96
69	0	48	0,00	97,96
70	1	49	2,04	100,00

19.2. Percentiles poblacionales de peso y talla en niños

Los percentiles utilizados habitualmente para evaluar el crecimiento de un niño, son estimaciones de los verdaderos percentiles poblacionales. Suelen obtenerse para el peso, la talla, el perímetro cefálico y el índice de masa corporal. Pueden hallarse las tablas actualizadas para Argentina en: http://www.garrahan.gov.ar/docs/2270/rgenerales.html

Los percentiles del **peso y talla** son los más utilizados. Se representan con curvas mostrando los percentiles 3, 10, 25, 50, 75, 90 y 97 en función de la edad, correspondientes a valores de niños normales, sanos.

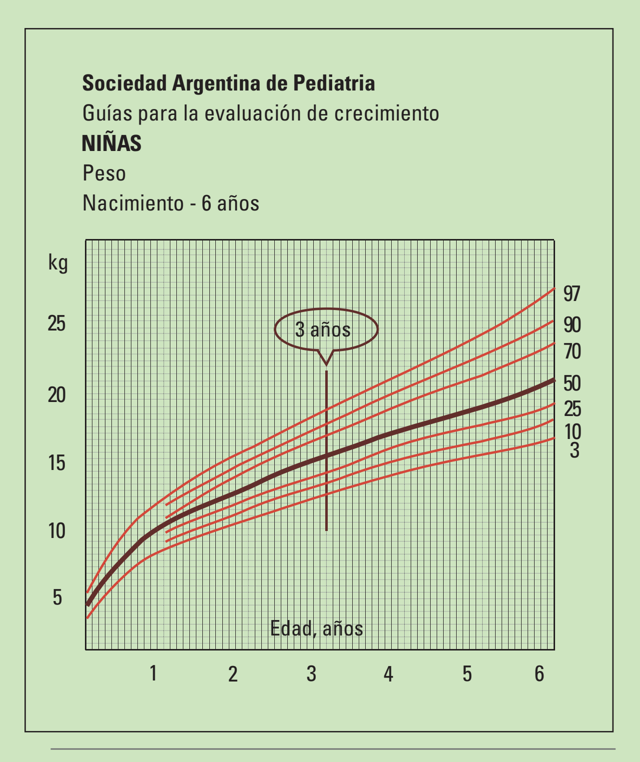


Figura19.1. *Fuente: Sociedad Argentina de Pediatría, Guías para la Evaluación del Crecimiento, 2001.*

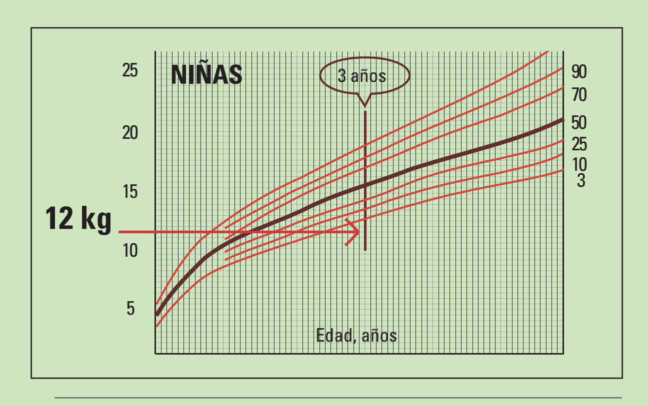


Figura19.2. *Peso de una niña de 3 años que se encuentra en el percentil 10.*

Ejemplo1:
La figura 19.1 muestra los percentiles de peso, desde el nacimiento hasta 6 años de edad.

Las curvas muestran los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de niñas sanas con edades entre 0 y 6 años. Se destacan los valores para 3 años de edad.

Si el médico dice que una niña de 3 años está en el percentil 10 respecto al peso significa que el 10 % de las niñas sanas de 3 años pesan a lo sumo como ella. Pero, ¿cuánto pesa? Como está en el percentil 10 y tiene 3 años podemos hallar su peso. Lo obtenemos (12 kg) trazando una línea horizontal, en el gráfico de los percentiles del peso, a la altura en que el percentil 10 corta la línea de 3 años (figura 19.2)

Ejemplo2:
La figura 19.3 (Sociedad Argentina de Pediatría, Guías para la Evaluación del Crecimiento, 2001) muestra los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de niñas sanas con edades entre 0 y 19 años. Este gráfico permite hallar el rango de valores de los pesos del 94% de las niñas sanas para cada edad. Solamente un 6% de niñas sanas tendrán pesos fuera de ese rango con un 3% por debajo y un 3% por encima.

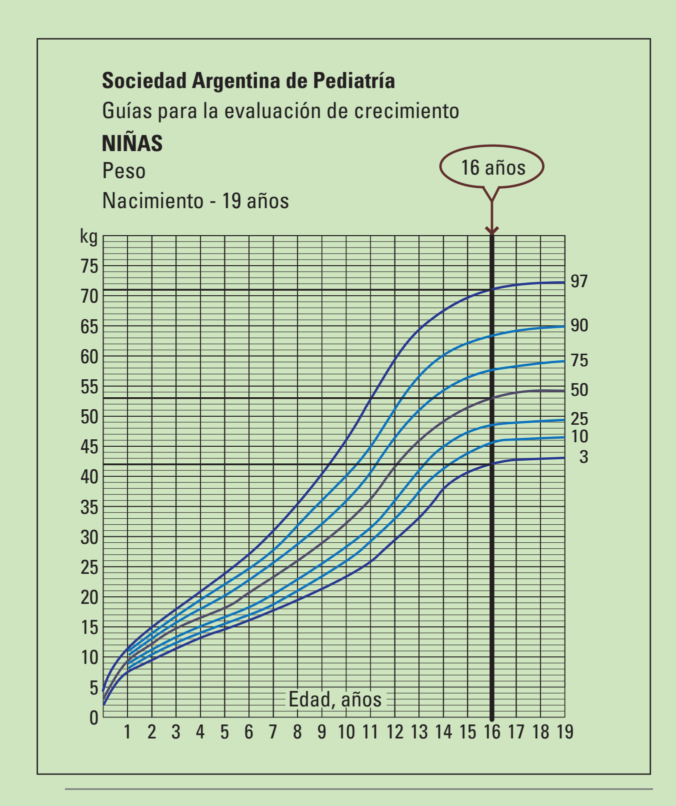


Figura19.3. *Las curvas muestran los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de niñas sanas con edades entre 0 y 19 años. Se destacan los valores para 16 años de edad.*

Percentiles del peso (kg) de las mujeres de 16 años. Tabla 19.3

Percentil	Peso
3	42,0
10	45,5
25	48,5
50	53,0
75	57,5
90	63,5
97	71,0

Fijemos nuestra atención en edad =16 años (figura 19.3). Se observa una línea vertical para esa edad. Los pesos que se obtienen, de los puntos donde la línea vertical corta a cada una de las curvas, se muestran en la tabla 19.3 y en la figura 19.4.

El rango de valores para el 94 % central de los datos se encuentra entre los pesos correspondientes a los puntos donde esa línea corta los percentiles 3 y 97 respectivamente. Este rango va desde 42 kg (percentil 3) hasta 71 kg (percentil 97). Solamente el 6% tiene su peso fuera de ese rango de valores, se trata de las extremadamente livianas y las extremadamente pesadas. La mediana es de 53 kg.

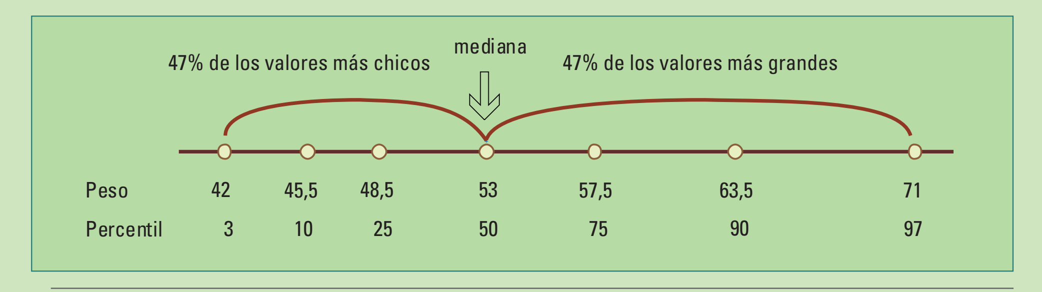


__Figura 19.4.__ Diagrama de puntos de los pesos correspondientes a los percentiles 3, 10, 25, 50, 75, 90 y 97 de las mujeres de 16 años.

En la figura 19.4 se aprecia la mayor concentración de los valores más chicos, por debajo de la mediana, en comparación con los valores más grandes. Por lo tanto el **peso es una variable no simétrica,** con leve asimetría hacia la derecha.

Ejemplo 3:
Pesa 46 kg y tiene 16 años. En relación con sus compañeras está en el percentil 20 (tabla 19.2), ¿y respecto de la población? Puede obtenerse esa respuesta utilizando los percentiles poblacionales de la figura 19.5. Un poco más del 10 % de las chicas de su edad pesan menos que 46 kg; está levemente por encima del percentil 10 y por debajo del percentil 25 respecto al peso.

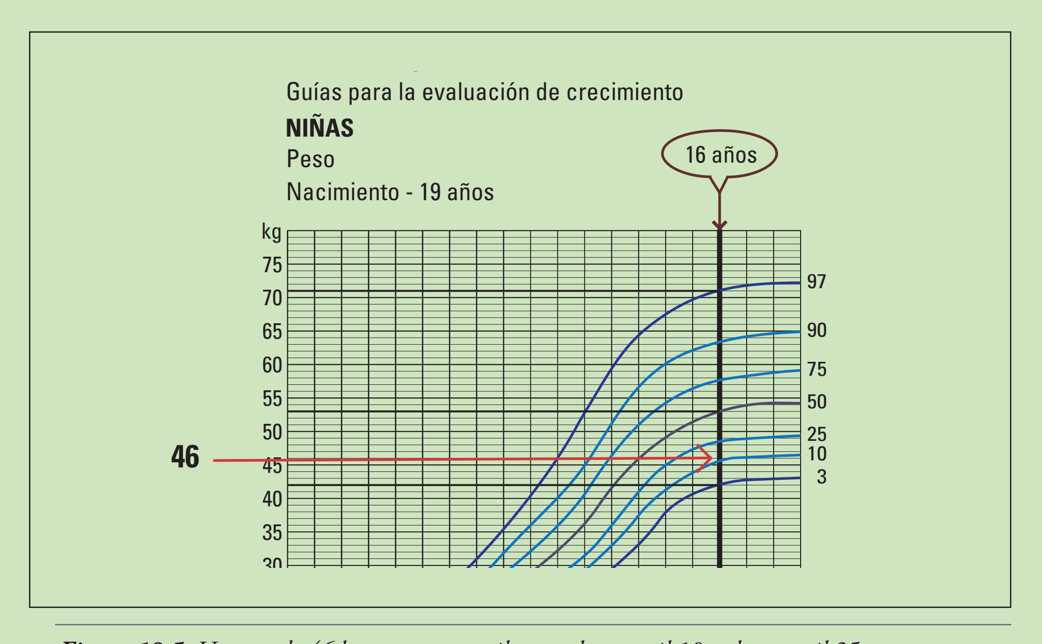


__Figura 19.5.__ Un peso de 46 kg, es un percentil entre el percentil 10 y el percentil 25.

19.3. Actividades y ejercicios

1. La mediana deja la mitad de los datos ordenados a cada lado.

- ¿Por qué en la figura 19.4 aparecen solamente el 47% de los valores más chicos a la izquierda de la mediana, el 47 % a la derecha y no el 50% a cada lado?
- ¿Faltan datos? ¿Cuál es el porcentaje de datos faltantes? ¿Por donde estarán?

2. Complete la tabla de frecuencias siguiente utilizando la información de la figura 19.4

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %	Frecuencia Relativa en % / Longitud del Intervalo
[42 ; 45,5)	3,5	10- 3 = 7	7/3,5 = 2
[45,5 ; 48,5)	3,0	25-10 = 15	15/3,0 = 5
[48,5 ; 53)	4,5	50-25 = 25	
[53 ; 57,5)	4,5	75-50 =	
[57,5 ; 63,5)			
[63,5 ; 71)			

3. Construya un histograma para el 94% central de los valores de la variable peso de las mujeres de 16 años utilizando la tabla anterior. Grafique en el eje horizontal los intervalos y utilice la escala densidad para el vertical (la frecuencia relativa en %) / (la longitud del intervalo). Es necesario utilizar la escala densidad porque los intervalos de clase del histograma tienen distinta longitud. Indique donde se encuentra el percentil 50 y observe que los datos presentan una leve asimetría a derecha.

4. Construya un histograma para el peso de las jóvenes de 13 años siguiendo los siguientes pasos:

a) Trace una línea vertical en la figura 19.3 en edad = 13 años.
b) Halle los pesos que corresponden a los puntos donde la línea vertical corta a cada una de las curvas. Esos pesos son los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso de chicas de 13 años.
c) Obtenga una tabla de frecuencias similar a la presentada en 2. pero en este caso con edad=13 años.
d) Construya un histograma para el 94% central de los valores de la variable peso de las chicas de 13 en forma similar a lo realizado en 3.

5. Utilice los pesos de los alumnos y alumnas de su división (1. de 18.4 Actividades y Ejercicios) para obtener tablas de porcentajes acumulados de los pesos de varones y mujeres por separado. Construya diagramas de tallo y hojas de los pesos de alumnos y alumnas por separado. Obtenga los percentiles del 3, 10, 25, 50, 75, 90 y 97 %.

6. Utilice los pesos de los alumnos y alumnas de los otros dos años (2. de 18.4 Actividades y Ejercicios) para obtener tablas de porcentajes acumulados de los pesos de varones y mujeres por separado. Construya diagramas de tallo y hojas de los pesos de alumnos y alumnas por separado. Obtenga los percentiles del 3, 10, 25, 50, 75, 90 y 97 %.

7. Grafique los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso para las mujeres obtenidos para cada una de las divisiones en un único gráfico, en función de la edad, de la siguiente manera:

   En el eje horizontal la mediana de la edad de las mujeres de cada año.
   En el eje vertical los percentiles del peso.

   Una los tres puntos de cada percentil.

8. Grafique los percentiles 3, 10, 25, 50, 75, 90 y 97 del peso para los varones obtenidos para cada una de los años en un único gráfico, en función de la edad, en forma similar al punto anterior. Una los tres puntos de cada percentil.

20. Curvas de densidad

Poderosa herramienta para describir la distribución de los datos.

Hemos desarrollado un conjunto de herramientas para describir la distribución de los datos: tablas de frecuencias, histogramas, diagramas tallo-hoja, cálculo de medidas resumen (media, mediana, desvío estándar, distancia intercuartil, percentiles), gráfico de caja y brazos. Algunas veces estas herramientas tienen inconvenientes:

- un diagrama tallo-hoja no es práctico para conjuntos con muchos datos.
- las tablas de frecuencias, así como sus representaciones gráficas (los histogramas), eliminan los detalles y dependen de la longitud de los intervalos de clase.
- las medidas resumen (media, mediana, desvío estándar, distancia intercuartil, percentiles) muestran aspectos parciales de los datos.

¿Es posible describir la distribución de los datos en forma completa mediante una única expresión?

La respuesta es: ¡Depende!

¿De qué depende?

Si estamos dispuestos a **describir el patrón general de los datos,** omitiendo los atípicos, **la respuesta es sí.**

Esa respuesta la provee la expresión de una curva - **un modelo matemático, curva de densidad** - para la distribución de los datos.

En la sección 17.2 presentamos algunos patrones especiales que pueden presentar los histogramas mediante curvas. Las expresiones de dichas curvas son precisamente los modelos que necesitamos. Se trata de **descripciones matemáticas idealizadas;** constituyen poderosas herramientas para describir la distribución de los datos. Son especialmente útiles cuando de trata de describir una cantidad muy grande de observaciones.

Podemos establecer un paralelo con la física del movimiento de los cuerpos. La ecuación de la recta describe un movimiento rectilíneo uniforme; pero, ningún desplazamiento real será perfectamente rectilíneo y uniforme. Si graficamos distancia en función del tiempo, con valores medidos de un desplazamiento real, los puntos no caerán exactamente sobre una recta, pero la recta es una buena descripción del movimiento cuando la velocidad es pareja y el desplazamiento es en una única dirección.

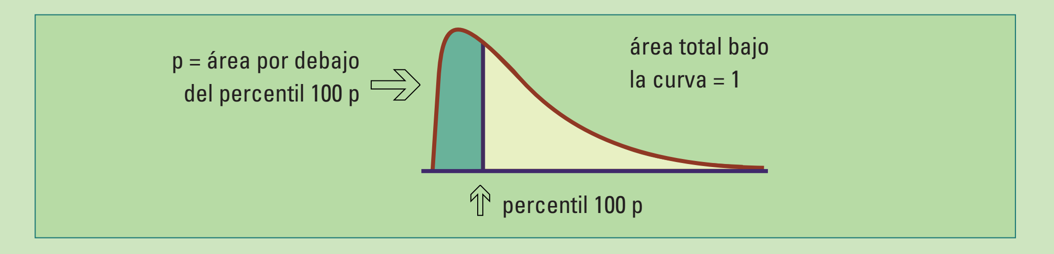
De la misma manera, como la recta es una de las muchas curvas requeridas para describir el desplazamiento de un objeto en función del tiempo, la Curva de Gauss o curva Normal es uno de los tipos de curvas que pueden utilizarse para describir los diferentes tipos de variabilidad de los datos.

□ 20.1. Medias resumen en curvas de densidad

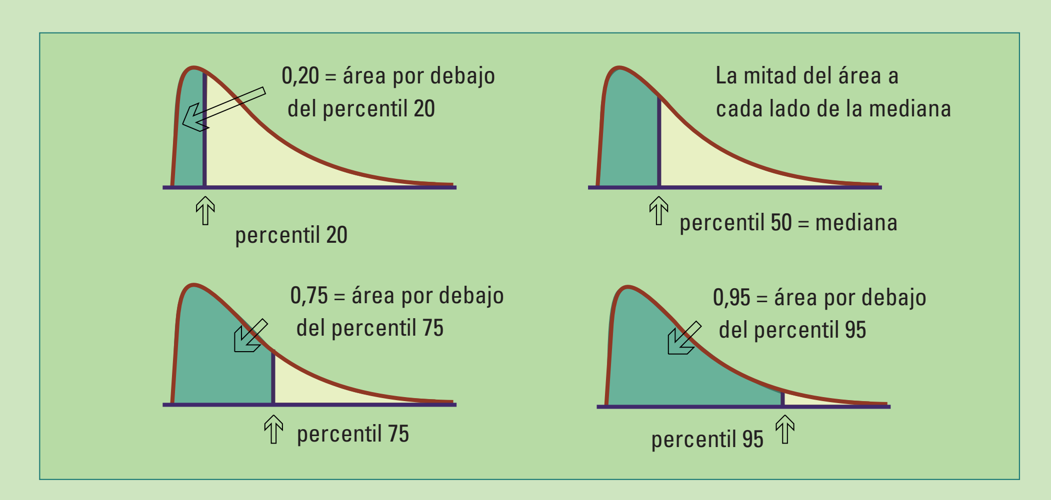
Las curvas de densidad son histogramas idealizados. Las medidas de posición y dispersión se aplican tanto a curvas de densidad como a conjuntos de datos.

Consideremos los percentiles en primer término.

Sabemos que una proporción p de observaciones está por debajo del percentil 100 p.



El percentil 100 x p de una curva de densidad es el punto sobre el eje horizontal para el cual queda a su izquierda el 100 x p % del área bajo la curva, o una proporción p.



En una curva de **densidad simétrica** es fácil **ver "a ojo" donde se encuentra la mediana,** el punto que divide al área en dos partes iguales.

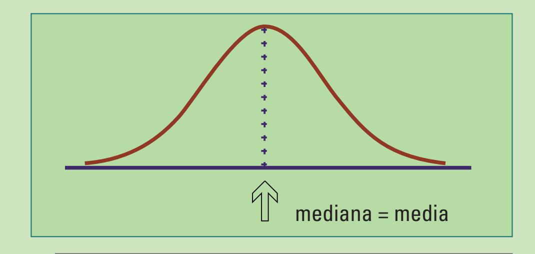


Figura 20.3. *Media y mediana en una curva de densidad simétrica.*

Para una **curva simétrica, la media, el punto de equilibrio coincide con la mediana** que divide el área en dos partes iguales (figura 20.3.).

Como parte de la idealización inherente a un modelo matemático, las curvas de densidad simétricas son "perfectamente simétricas" aunque los datos reales rara vez presenten una simetría perfecta.

Para cualquier curva general **no es fácil hallar a ojo** la mediana, la media y los percentiles. Pero es posible utilizar integrales para obtenerlos. Las integrales son herramientas de análisis matemático que permiten obtener el área por debajo de una curva cuando se conoce la expresión de la misma. No lo haremos aquí.

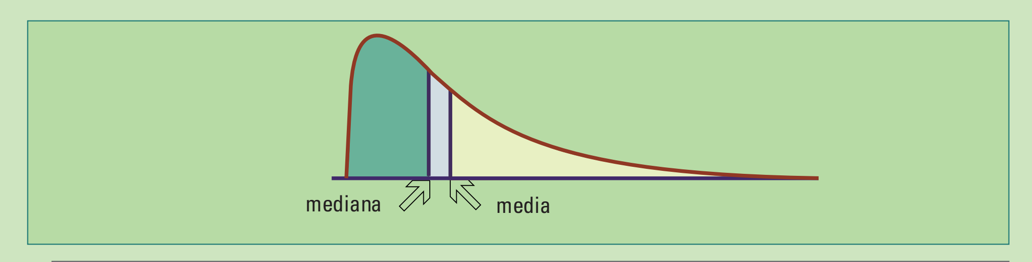


Figura 20.4. *Media y mediana en una curva de densidad asimétrica a derecha.*

La media es el punto de equilibrio de una vara sin peso sobre la que se colocan en cada punto correspondiente al valor de cada dato, pesos idénticos (sección 18.1.1.). La **vara no queda en equilibro** si se apoya en cualquier otro punto. Esta interpretación se extiende a curvas de densidad.

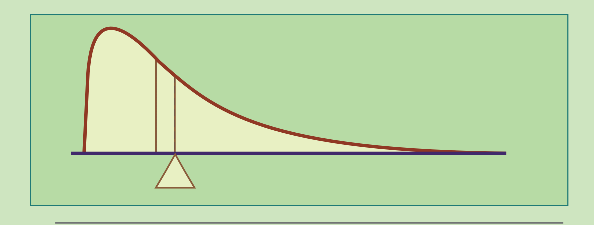


Figura 20.5. *La media es el punto donde la curva de densidad quedaría en equilibrio.*

En una curva asimétrica la media (el punto de equilibrio) es arrastrado hacia la cola larga de la distribución más que la mediana (figuras 20.4 y 20.5). Hallar a ojo la media en una curva asimétrica es más difícil que la mediana, pero la podemos obtener mediante integrales (no lo haremos aquí).

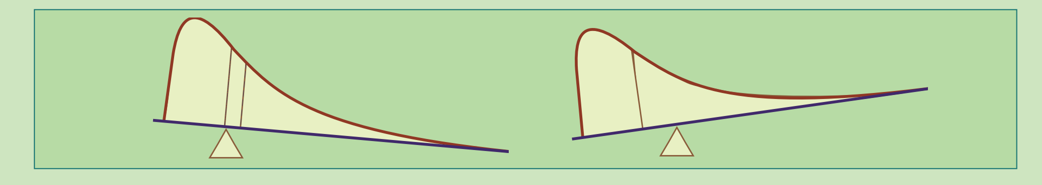


__Figura 20.6.__ La curva no queda en equilibrio cuando se apoya en un punto diferente de la media.

La parte inferior de la figura 20.6 ilustra que **la curva no queda en equilibrio cuando se apoya en la mediana.** El área bajo la curva del lado derecho de la mediana “pesa más”. Decimos que la distribución tiene **cola pesada a derecha.**

> **Media y mediana de una curva de densidad:** La mediana es el punto que divide el área bajo la curva en dos partes iguales. La media es el punto de equilibrio o centro de gravedad, sobre el cual quedaría en equilibrio si se construyera con un material sólido.

Para calcular la mediana a ojo tratamos de dividir el área en dos partes iguales. Para hallar los **cuartiles,** tratamos de dividir el área por debajo de la curva de densidad en 4 partes iguales (figura 20.7).

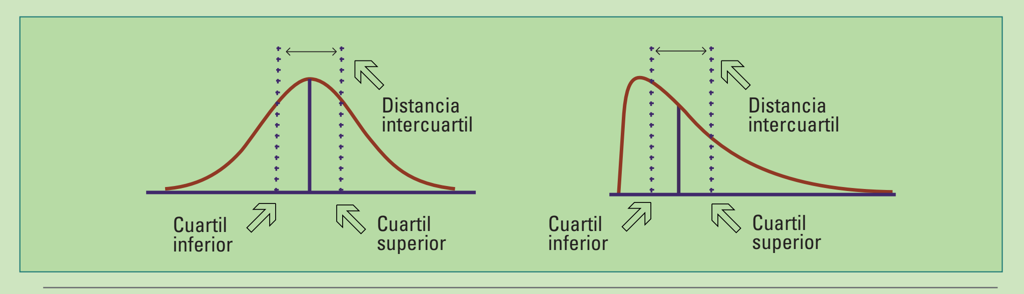


__Figura 20.7.__ Los cuartiles, la mediana y la distancia intercuartil en una curva simétrica y en una curva asimétrica a derecha.

La distancia intercuartil es la diferencia entre el cuartil superior y el inferior (también llamados tercer y primer cuartil).

Los cuartiles, por lo tanto, la mediana y la distancia intercuartil, pueden calcularse en forma aproximada a ojo para cualquier curva de densidad. Esto no ocurre con el desvío estándar (18.2.3), que no es una medida natural para la mayoría de las distribuciones. Cuando es necesario, el desvío estándar correspondiente a una curva de densidad, también (como dijimos para los percentiles), puede calcularse utilizando integrales. No se desarrollará esta forma de calcular en este libro.

La curva de densidad es una descripción idealizada de la distribución de los datos, por eso distinguimos la **media** y el **desvío estándar** de una **curva de densidad** de los números

y s (media muestral y desvío estándar muestral respectivamente) y se obtienen a partir de un conjunto de datos. La forma habitual de indicar la media de una distribución idealizada es mediante la letra griega "mu": **μ.** El desvío estándar se indica por **σ**, la letra griega "sigma".

20.2. Ventajas de la curva Normal

¿Para qué sirve tener un conjunto de datos cuyo histograma es aproximadamente Normal? ¿Por qué se habrá enamorado Galton de la curva gaussiana? (sección 17.1)

Hemos visto que es muy útil reemplazar un conjunto de datos por unos pocos valores, las medidas resumen, para describir sus características generales.

> Cuando los datos tienen una **distribución Normal** la distribución de los mismos se puede reducir a **dos números:** la media y el desvío.

En general, es deseable tener **patrones** que representen **la forma** de la distribución de **los datos** y que permitan además representar sus características más importantes mediante **una cantidad pequeña de números.**

20.2.1. Histogramas y la curva Normal

Pensemos primero en un conjunto con **muchísimos datos.** Podemos construir histogramas con intervalos de distinta longitud y superponerle una Curva de Gauss. Como los datos son muchísimos podemos achicar la longitud de los intervalos de clase tanto como queramos.

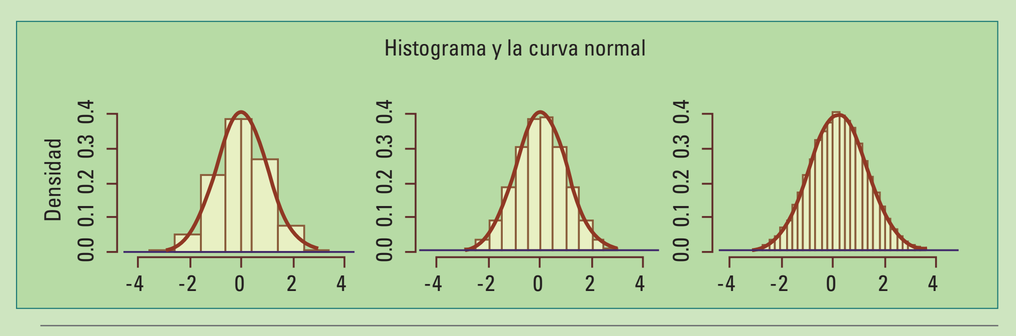


Figura 20.8. *Superposición de la curva de Gauss a histogramas con intervalos de longitud decreciente.*

A medida que se achica la longitud de los intervalos de clase mejora la aproximación de la campana de Gauss.

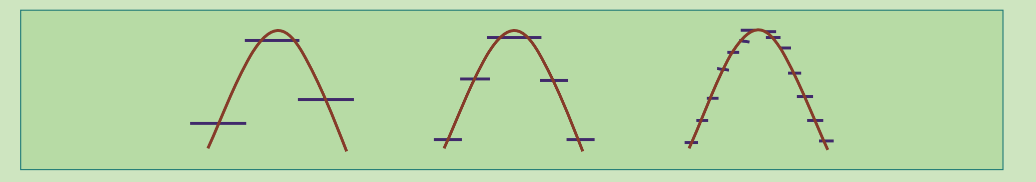


Figura 20.9. *Detalle, en un sector ampliado, de la arista superior de los rectángulos de clase de los histogramas y su aproximación creciente a la curva de Gauss.*

El primero de los histogramas muestra escalones resultantes del agrupamiento de los datos en intervalos de clase, pero estas irregularidades disminuyen al reducir la longitud de los intervalos (figura 20.9). La curva de Gauss en la figura 20.8 describe la distribución de los datos en forma más precisa que los histogramas.

Cuando un histograma se grafica utilizando las frecuencias en el eje vertical, la escala depende de la cantidad de datos. Si se utilizan frecuencias relativas o porcentajes esto es menos arbitrario y el **área del rectángulo es proporcional** a la frecuencia relativa.

Es más natural que el **área del rectángulo** sea **igual a la frecuencia relativa.** Lo logramos si en el **eje vertical** graficamos la **frecuencia relativa dividida la longitud del intervalo.** Esto se llama **escala de densidad** y permite tener la misma escala vertical aunque cambiemos la longitud de los intervalos y el área total de los rectángulos del histograma siempre 1 (ó 100 si las frecuencias relativas están expresadas como porcentajes).

La curva que describe la forma de la distribución se llama **curva de densidad** y tiene área 1. El **área bajo la curva** sobre cualquier **intervalo de valores** del eje horizontal es **la proporción de observaciones** que caen en ese intervalo.

En la figura 20.8 la escala de densidad va de 0 a 0,4 en los 3 histogramas y en la curva. Podemos calcular en forma aproximada el área total pues la figura es aproximadamente un triángulo cuya base tiene longitud aprox. 5 y la altura es aprox. 0,4. El cálculo aproximado resulta:

$$\frac{\text{long de la base x altura}}{2} = \frac{5 \text{ x } 0{,}4}{2}$$

$$\frac{\text{long de la base x altura}}{2} = 1$$

20.2.2. Media y desvío de la curva normal

Todas las curvas Normales son simétricas, tienen un único pico y forma de campana.

Sus colas caen rápidamente, por lo tanto no se esperan valores muy alejados (outliers). La media, la mediana y el pico coinciden en el centro de la curva.

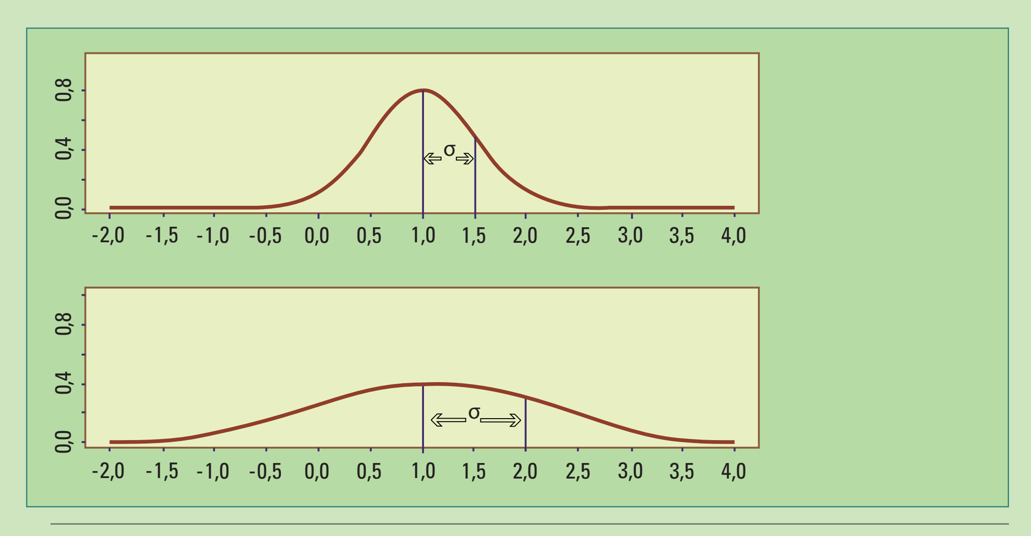


Figura 20.10. *Dos curvas normales con media 1 y distintos desvíos.*

Otra propiedad importante de la curva de densidad Normal es poder localizar el desvío estándar a ojo: a medida que nos movemos en ambas direcciones desde el centro **μ** de la curva, ésta aumenta su pendiente

hasta un punto (punto de inflexión), a partir de allí la pendiente empieza a disminuir



Los dos puntos en los cuales ocurre este cambio de curvatura están localizados a una distancia **σ** a cada lado del centro **μ**.

μ es la media
σ es el desvío

Recuerde, **μ** y **σ** solos no determinan la forma de una distribución en general. Éstas son propiedades de las distribuciones gaussianas.

Pero...

> Cuando los valores de una variable tienen distribución Normal, **sólo dos números** alcanzan para determinar la distribución de todos sus valores. Esos dos números, **μ** y **σ** son **los parámetros** de la distribución Normal.

Pero...

> Pequeños alejamientos de la distribución Normal pueden llevar a que **μ** y **σ** no signifiquen nada.

Un detalle extra:

Siempre es más seguro utilizar los percentiles porque tienen el mismo significado en todo tipo de distribuciones. Cuando no hay grupos aislados, las 5 medidas resumen: mínimo, cuartil inferior, mediana, cuartil superior y máximo, son en general una buena representación de los datos.

20.2.3. Otras característcas interesantes

Si un histograma se aproxima por una curva Normal podremos decir algunas cosas más que, simplemente, caracterizar su media y su desvío.

Podremos establecer **criterios** sobre **donde se encuentra** la mayoría de los **datos.**

Los **criterios** que veremos a continuación se utilizan cuando **podemos suponer** que los **datos** tienen una distribución **aproximadamente Normal,** por la naturaleza del experimento, con **μ** y **σ** conocidos. Cuando no son conocidos se estiman mediante la media muestral (x) y el desvío estándar muestral (s), respectivamente, tal como vimos en las secciones 18.1.1 y 18.2.3:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{y} \qquad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

donde los x_1 son los datos y n es la cantidad total ($x_{1,}$ $x_{2,}$ ... x_n)

Utilizaremos estas reglas en el próximo capítulo sobre control de calidad.

Si una distribución tiene una forma gaussiana, entonces vale la siguiente **regla 68-95-99,7 :**

- Aproximadamente el 68% de los valores se encuentran dentro de 1 desvío estándar (**σ**) de la media (**μ**). Es decir que bastante más de la mitad de los valores están comprendidos dentro del intervalo (μ - **σ**, μ +**σ**) ó μ ± **σ** (figura 20.10).
- Cerca del 95% de los valores están entre la media menos 2 veces el desvío estándar y la media más 2 veces el desvío estándar, o sea dentro de**σ**μ ± 2 (figura 20.11).

- El 99,7% (casi todos) de los valores se encuentran en el intervalo (**μ** - 3**σ**, **μ** + 3**σ**), o sea dentro de μ ± 3 **σ** (figura 20.12).

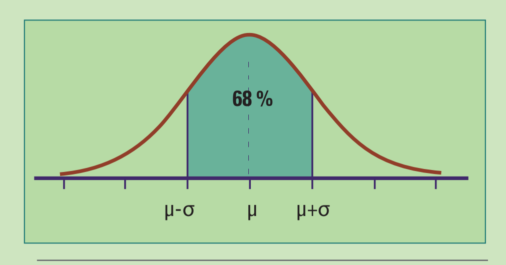


Figura 20.11.

Como la mayoría de los valores en una distribución Normal se encuentran en la zona central, alrededor de la media (**μ**), el 68% de los valores están a una distancia no mayor al desvío. Al alejarnos un desvío más de la media, hacia los dos lados, agregamos más valores (un 14% a cada lado); pero, son menos porque se trata de una zona de menor concentración de datos. Obtenemos así el intervalo (**μ**-2**σ**, **μ**+2**σ**) allí se encuentra aproximadamente el 95% de los valores.

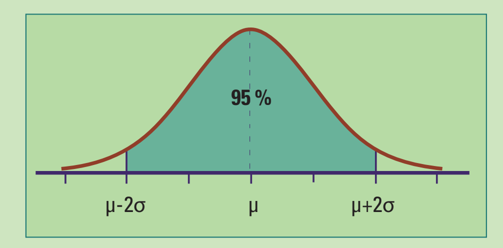


Figura 20.12.

Alejándonos otro desvío más, agregamos apenas un 2% de cada lado, llegando a 99.7 % en el intervalo (**μ**-3 **σ**, μ+3**σ**).

Ejemplo: Un taller metalúrgico produce remaches cuya longitud debe ser de 5 cm con una tolerancia de 0,3 cm (5 ± 0,3 cm). Por lo tanto, las longitudes aceptables están en el intervalo (4,7; 5,3). Interesa evaluar la calidad de la producción teniendo en cuenta este requerimiento.

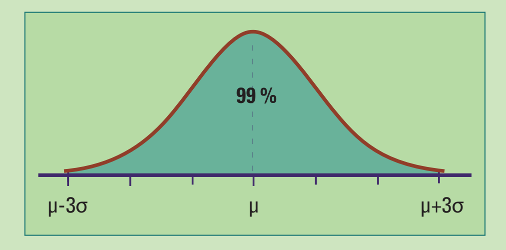


Figura 20.13.

Como suele ocurrir en esta industria, si la producción se realiza en condiciones normales tendremos muchos remaches cuya longitud esté cerca de 5 cm y pocos alejados; las longitudes tendrán una distribución gaussiana.

Supongamos que los registros históricos de la producción de estos remaches, con el mismo equipamiento, muestran que la media de las longitudes es efectivamente 5 cm con un desvío de 0,15 cm (**μ**=5 y **σ**=0,15).

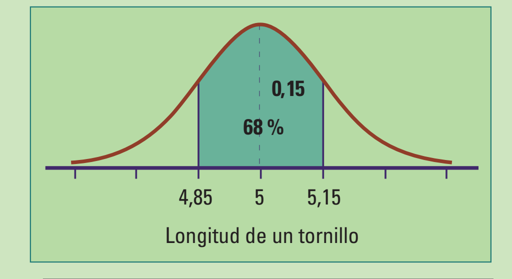
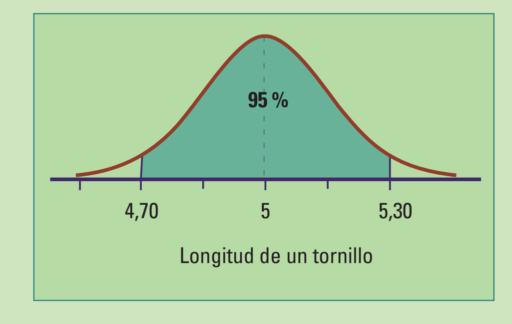


Figura 20.14

Luego:

- Aproximadamente el 68% de los valores se encuentran dentro de1 desvío estándar (σ) de la media (μ). Es decir que bastante más de la mitad de los valores están comprendidos dentro del intervalo (μ-σ ,μ+σ) ó μ ± σ (figura 20.11).
- Cerca del 95% de los valores están entre la media menos 2 veces el desvío estándar y la media más 2 veces el desvío estándar, o sea dentro de μ ± 2**σ** (figura 20.12).



- El 99,7% (casi todos) de los valores se encuentran en el intervalo (μ-3**σ** ,μ+3**σ**), o sea dentro de μ ± 3**σ** (figura 20.13).

Casi el 5% de los remaches tendrá una longitud por fuera de los límites especificados. El encargado de control sabrá si este porcentaje de remaches a desechar (scrap) es admisible. Si no lo es deberán modificarse los procesos de producción, hasta que se logre un perfil de calidad adecuado.

20.2.3.1. Regla 68 - 95 - 99,7

Supongamos que un conjunto de datos (x_1, x_2 ... x_n) tiene una distribución gaussiana con media **μ** y desvío estándar **σ**. El conjunto **de datos estandarizados** (z_1, z_2 ... z_n), o "puntajes z", que se obtiene restando **μ** y dividiendo por **σ** ($z_i = \frac{x_i - \mu}{\sigma}$), tendrá una distribución Normal Estándar (figura 20.15).

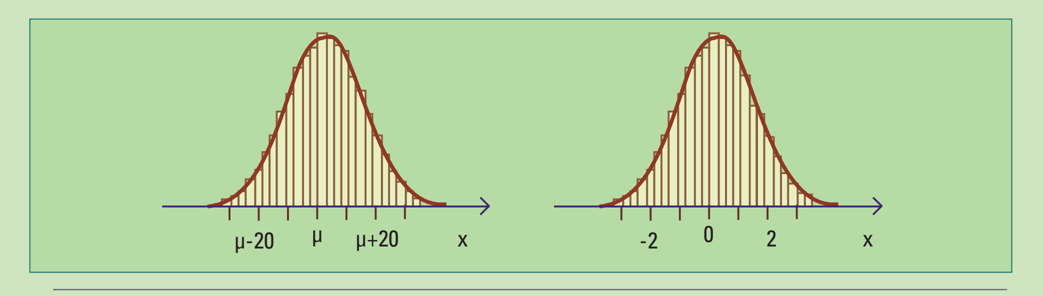


Figura 20.15. *Histogramas de un conjunto de datos en su escala original (x) y transformados en puntaje z.*

Recordemos (sección 17.1.1) que la curva Normal Estándar, también llamada N(0,1), está **dada por** $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ **, y no depende de parámetros desconocidos.**

Las áreas bajo esta curva se pueden calcular.

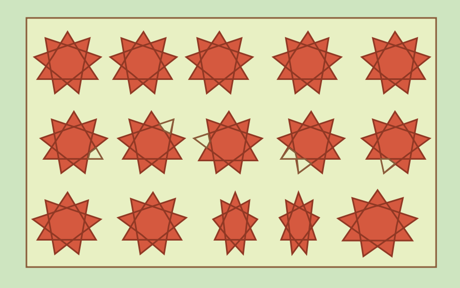
¡Conocer el área bajo la curva Normal Estándar sobre cualquier intervalo, permite conocerla para todos los intervalos bajo cualquier curva Normal!

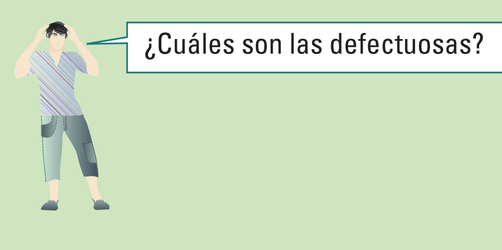
En particular, las áreas sobre los intervalos, (**μ-σ**; **μ+σ**); (**μ-2σ**; **μ+2σ**) y (**μ-3σ**; **μ+3σ**) bajo la curva N (**μ**, **σ**) son iguales a las áreas sobre los intervalos (-1,1) ; (-2,2) y (-3;3) bajo la curva N(0,1).

Los valores 68; 95 y 99,7 para los porcentajes de áreas por encima de los intervalos (-1,1); (-2,2) y (-3;3) bajo la curva Normal Estándar son aproximados. Valores más precisos son: 68,27; 95,45 y 99,73 respectivamente.

21. Control de calidad

Idealmente todas las unidades fabricadas serán idénticas y perfectas.





¿Para qué se realiza el control de calidad?

El control estadístico de calidad de productos y servicios tiene como objetivo **reducir su variabilidad** e idealmente eliminar sus defectos. Como un ideal, se busca que todas las unidades fabricadas sean idénticas y perfectas. En la práctica real se consigue reducir los desperdicios, minimizar los reprocesamientos y mejorar la opinión del cliente, tratando de hacer las cosas bien de una primera vez.

En el desarrollo de un producto puede servir en experimentos para comparar materiales, componentes o ingredientes. Durante el proceso de producción y distribución, los métodos estadísticos permiten identificar problemas que atentan contra la calidad buscada.

Hemos visto (secciones 17.1.1 y 20.2.3) que diferentes piezas, correspondientes a un mismo producto, pueden parecer iguales pero al medir sus características detalladamente se encuentran diferencias. Por más cuidado que se tenga en la calibración de las máquinas, se controlen los factores ambientales, se vigilen los materiales y se capaciten los operarios, **las piezas no serán idénticas.** Se trata de una **variabilidad** natural o **aleatoria.** Puede considerarse como un "ruido de fondo" inevitable.

Cuando el ruido de fondo de un proceso de producción es relativamente pequeño se lo considera aceptable. Cuando un proceso sólo está afectado por esa variabilidad aleatoria decimos que se trata de un "sistema estable" y "bajo control estadístico" o simplemente "en control".

Walter A. Shewhart identificó las variaciones que se presentan cuando el proceso productivo opera normalmente. Son el resultado de muchas causas generalmente pequeñas e inevitables, que ocurren todo el tiempo. Las llamó variaciones "debidas a **causas comunes"**, en contraposición con un segundo tipo de variaciones, las debidas a **"causas especiales o asignables"** y que ocurren de vez en cuando.



Walter Andrew Shewhart. (1891-1967). Físico, matemático y estadístico norteamericano, también conocido como el padre del control estadístico de la calidad.

No es posible –ni tiene sentido- perder tiempo en averiguar la causa de una variación debida a causas comunes cuando el proceso ya satisface las especificaciones. Sí es útil dedicarse a las debidas a causas especiales; usualmente provienen de tres fuentes: de los equipos, del operador o de las materias primas utilizadas.

Por ejemplo, se habla de causas especiales cuando la calidad del producto es afectada por haberse utilizado materias primas defectuosas o por el accionar inadecuado de los operarios que, por cansancio ó distracción, en muchas oportunidades continúan la producción sin advertir un desajuste en su máquina.

No son causas comunes. Se trata de **causas asignables:** máquinas mal ajustadas, materias primas defectuosas, errores del operador, software incorrecto, etc. Las piezas producidas bajo estas condiciones anómalas no tendrán su variabilidad habitual. Esta variabilidad extra suele ser grande, en comparación con el ruido de fondo, y representa un nivel inaceptable del rendimiento del proceso.

Cuando un proceso de producción opera en presencia de **causas asignables** decimos que **está fuera de control.**

Al controlar un proceso interesa restringir la variación únicamente a la debida a las causas comunes; **las causas asignables deben ser detectadas y eliminadas.** Cuando **la única variación presente** es debida a **causas comunes,** y no a una causa asignable decimos que **el proceso opera en estado bajo control** o que **está en control.** Un proceso en control es **estable en el tiempo** respecto a las variaciones y no muestra indicaciones de causas extrañas.

Un proceso en control (o bajo control) es un proceso estable.

No significa necesariamente que se satisfagan las especificaciones del producto.

La variabilidad debida a causas comunes puede exceder los límites de tolerancia del producto. Puede ocurrir que la proporción de piezas defectuosas sea mayor a lo tolerable en términos económicos. En ese caso se deberán introducir modificaciones al proceso.

Cuando ya se ha establecido un proceso que cumple con las especificaciones de tolerancia del producto, el problema principal consiste en determinar cuándo el proceso está fuera de control.

Los gráficos de control, también llamados cartas de control de Shewhart, permiten reconocer situaciones en las cuales las causas asignables pueden estar afectando negativamente la calidad de un producto. Se trata de una secuencia de puntos obtenidos de muestras de piezas tomadas a través del tiempo. Son los valores de algún estadístico, tal como la media de la longitud o la proporción de piezas defectuosas.

□ 21.1. Gráficos de control

Un gráfico de control muestra en el eje vertical los valores de los datos y en el horizontal el orden en que fueron recogidos a través del tiempo. Es esencial, para este tipo de

gráfico que las muestras hayan sido tomadas en forma sucesiva en el tiempo, pues es esa evolución -de alguna o algunas características de un producto- lo que interesa controlar.

Tiene una línea horizontal a la altura de un **valor central,** ésta puede provenir de las especificaciones del producto o de valores históricos y dos líneas -una en el **límite inferior de control** (LIC) y otra en el **límite superior de control** (LSC)- para indicar cuan lejos por encima o por debajo del valor objetivo se espera que se obtengan los valores de las muestras.

Se grafican pesos, volúmenes, cantidades o más frecuentemente pesos promedio, el promedio de volúmenes o cantidades promedio (proporciones). Si los puntos caen dentro de los límites inferior y superior se considera que el proceso se encuentra en el estado de control estadístico.

Ejemplo: Las bolsas de galletitas siempre parecen tener menos unidades que las que debieran. Supongamos que un fabricante está llenando bolsas con galletitas de agua, y el valor objetivo es de 50 piezas por bolsa, con LIC = 45 piezas y LSC = 55 piezas. Supongamos, además, que 8 bolsas de galletitas son seleccionadas mediante un muestreo sistemático. Una de cada 200 bolsas, de una línea de producción, son inspeccionadas obteniéndose los siguientes resultados: 49 galletitas, 47 galletitas, 51 galletitas, 49 galletitas, 46 galletitas, 53 galletitas, 48 galletitas, y 55 galletitas.

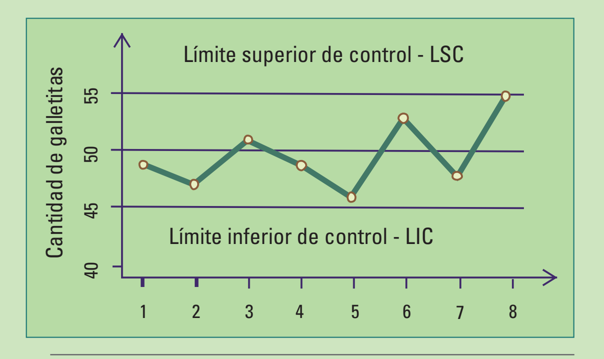


Fig. 21.1. Gráfico de control para la cantidad de galletitas.

El proceso que muestra la figura 21.1. parece estar operando en control, al menos por el momento.

> Un gráfico de control puede indicar una condición fuera de control cuando uno o más puntos caen más allá de los límites o presenta algún patrón de comportamiento no aleatorio.

21.1.1. ¿Cómo se establecen el valor central y los límites en un gráfico de control?

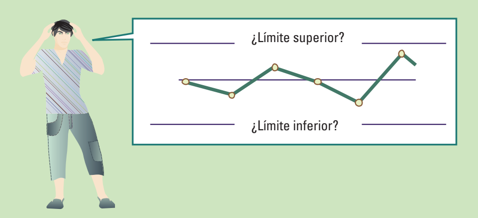
La idea detrás de los gráficos de control es obtener límites, fijar cotas para toda la variabilidad aleatoria, excluyendo la variabilidad asignable no deseada. De esta forma, las causas asignables tenderán a dar valores fuera de los límites de control mientras que la variabilidad aleatoria tenderá a generar puntos que se encuentran dentro de esos límites.

En general el valor central o **valor objetivo** es establecido por las especificaciones del producto (50 en el ejemplo de las galletitas). Otras veces es el promedio de muchos valores

(datos históricos) de la característica de interés, obtenidos con el proceso en estado de control. Por ejemplo, en una producción de almohadones el valor central podría ser su peso promedio cuando la variabilidad presente se debió únicamente a causas comunes.

Las especificaciones de los **límites de control** dependen de los límites de tolerancia del producto y de qué proporción de artículos está dispuesto a perder el fabricante.

Supongamos que el fabricante puede tener hasta un **5% de artículos** fuera de las especificaciones. Para el caso del llenado de bolsas de galletitas, esto significa que el 95% de las bolsas deben contener entre 45 y 55 galletitas. Si la distribución de la cantidad de galletitas por bolsa puede aproximarse por la Normal (ya vimos que suele ocurrir), cerca del 95% de los valores se encontrará dentro de 2 desvíos de la media (sección 20.2.3). El intervalo de cantidades aceptables [45;55] tiene longitud 10 y debe corresponder a 2 desvíos para que se cumpla que el 95% de las bolsas contengan entre 45 y 55 galletitas (sección 20.2.3). Luego la cantidad de galletitas por bolsa producida puede tener un desvío menor o igual a 10/2=5.



Si ahora el fabricante es más exigente, admitiendo sólo un 0,3% de productos defectuosos, entonces el 99.7% de las bolsas deben contener entre 45 y 55 galletitas, el intervalo [45; 55] de longitud 10 debe corresponder a 3 desvíos y el desvío debe ser menor o igual a 10/ 3=3,33. El proceso de fabricación ahora requiere un desvío de 3,33 galletitas por bolsa.

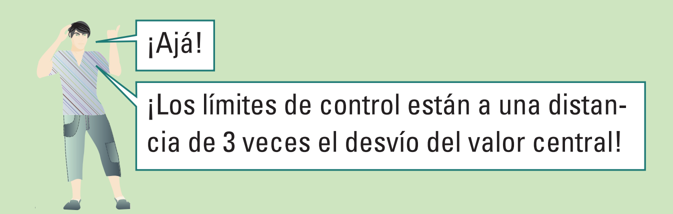
Una vez logrados los requerimientos de calidad se debe monitorear el proceso para garantizar que se mantenga estable.

21.1.1.1. Límites de control 3 - sigma

Es habitual colocar los límites de control a una distancia de más y menos 3 desvíos, a partir de la línea central, de la variable graficada. Estos límites se conocen como **límites de control 3 – sigma.**

- Límite inferior de control (LIC) = $\boldsymbol{\mu} - 3\boldsymbol{\sigma}$
- Límite superior de control (LSC) = $\boldsymbol{\mu} + 3\boldsymbol{\sigma}$

Si la variable graficada tiene distribución Normal con media igual al valor central del gráfico el **99,7% de los valores** (casi todos) **caerán dentro de los límites de control 3 – sigm**a (ver sección 20.2.3). Con estos límites tendremos un 0,3 % de falsas alarmas.



Ejemplo: Una fábrica produce garrafas de gas comprimido de uso doméstico con 20 kg de capacidad nominal. La capacidad volumétrica es una de las características importantes de las mismas. Cada hora se selecciona una garrafa de la línea de producción y se mide su capacidad volumétrica interna en dm^3. En una jornada de 16 horas (2 turnos) se obtuvieron los siguientes valores: 45,91; 46,34; 47,52; 46,52; 47,15; 47,15; 47,99; 46,81; 45,70; 47,25; 45,85; 48,14; 47,56; 48,01; 46,55; 47,27.

Se sabe que la capacidad volumétrica interna de una garrafa, cuando el proceso de producción opera en condiciones de control, es una variable con distribución Normal con media 47 dm^3 y desvío estándar de 0,666 dm^3.

> No confundir la capacidad nominal de 20 kg con la capacidad volumétrica que se mide en dm^3 y su valor está alrededor de 47.

La carta de control de $3 - \boldsymbol{\sigma}$ (3 desvíos) tendrá los siguientes límites:

- Límite inferior de control (LIC) = 47-3 x 0,666
  = 45,002
  = 45
- Límite superior de control (LSC) = 47+3 x 0,666
  = 48,998
  = 49

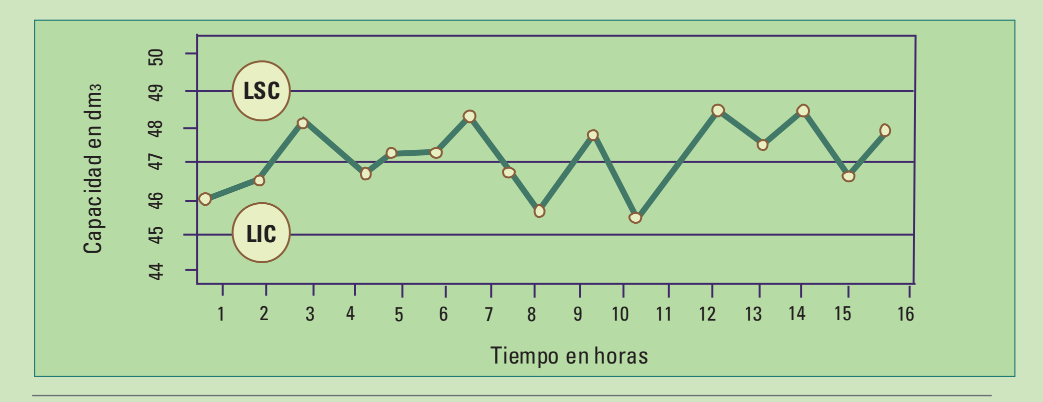


Figura 21.2. *Gráfico de control para la capacidad volumétrica de garrafas de uso doméstico.*

La figura 21.2 muestra los siguientes valores: 45,91; 46,34; 47,52; 46,52; 47,15; 47,15; 47,99; 46,81; 45,70; 47,25; 45,85; 48,14; 47,56; 48,01; 46,55; 47,27 de las capacidades volumétricas de 16 garrafas, seleccionadas una cada hora en una jornada laboral de 2 turnos. Los puntos se encuentran dentro de los límites de control; el proceso se encuentra bajo control estadístico.

21.1.1.2. Estimación de los parámetros del proceso

Cuando un gráfico de control muestra un proceso bajo control estadístico es posible utilizar los puntos del mismo para estimar la media (μ), el desvío ($\boldsymbol{\sigma}$) y la fracción que no cumple con los requerimientos. Esto permite tomar decisiones respecto a realizar o no modificaciones en el ciclo de producción y actualizar los límites de control si fuera necesario.

21.2. Gráficos de control $\overline{x}$ (equis barra)

La mayoría de los procesos productivos son monitoreados seleccionando muestras y calculando promedios. En los gráficos de control, en vez de valores individuales se grafican **promedios,** calculados a partir de los datos de subgrupos o **muestras** de artículos.

> Los gráficos de control que utilizan promedios se denominan "Gráficos de control equis barra".

21.2.1. ¿Cómo se eligen las muestras de artículos?

Los subgrupos de artículos se eligen, en lo posible, de manera que contengan la variabilidad natural del proceso y excluyan la variabilidad debida a causas asignables. Describiremos dos tipos de criterios para elegirlos.

21.2.1.1. Unidades cercanas en el tiempo



Figura 21.3. Esquema de muestreo de 3 unidades cercanas en el tiempo cada 10 producidas (período de muestreo). Los puntos sólidos representan las unidades muestreadas.

Cada subgrupo está formado por unidades producidas al mismo tiempo o casi al mismo tiempo. De esta manera se espera que un cambio en el proceso ocurra entre subgrupos y no dentro de cada subgrupo. Se obtiene así como una foto del proceso en cada instante en que se tomó la muestra. Este tipo de muestreo suele utilizarse para detectar corrimientos en la media del proceso.

La figura 21.3 muestra un esquema de **muestreo sistemático** de unidades cercanas en el tiempo. Se eligen 3 artículos sucesivos al comienzo de cada período de 10 unidades. Los puntos muestreados aparecen como puntos negros sólidos. Es mejor elegir al azar, dentro de cada período, el momento a partir del cual se seleccionan los 3 artículos sucesivos.

21.2.1.2. Unidades representativas de un período completo

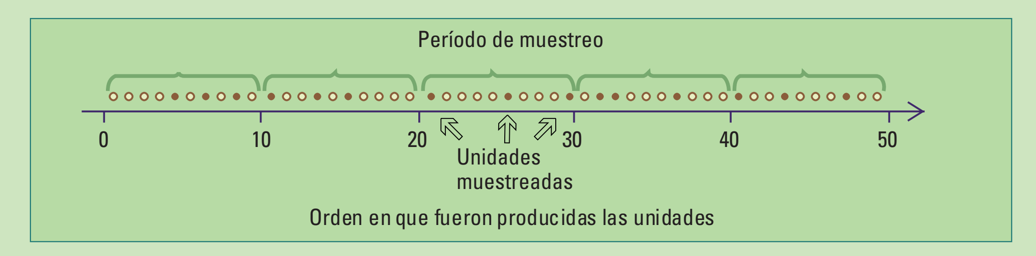


Figura 21.4. *Esquema de muestreo de 3 unidades representativas de todo el período de muestreo. Los puntos sólidos representan las unidades muestreadas.*

Cada muestra está formada por unidades representativas de todas las unidades producidas desde que se tomó la última muestra. Se trata de una muestra aleatoria de toda la salida del proceso sobre el intervalo de muestreo.

Este procedimiento suele utilizarse para tomar decisiones respecto de la aceptación o no de todas las unidades producidas en ese período.

La figura 21.4 muestra un ejemplo en el cual 3 unidades fueron elegidas al azar dentro de cada período de 10 unidades.

□ 21.2.2. ¿Cómo se calculan los límites de control tres sigma en un gráfico $\bar{X}$?

En un gráfico de control $\bar{X}$ (equis barra) **se grafican promedios** y para los límites de control se utilizan los **errores estándar.**

Un **error estándar** es el **desvío estándar** de las **medias muestrales.**

Esto significa que los límites de control 3 - sigma se establecerán como el valor objetivo más menos 3 veces el error estándar.

El error estándar se calcula como el desvío estándar (σ) dividido la raíz cuadrada de n, siendo n el tamaño de la muestra: Error estándar = $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

El error estándar siempre es menor que el desvío estándar. Esto es porque los promedios utilizan más información que un único dato y por lo tanto varían menos entre una muestra y la siguiente. Vimos un ejemplo de esa situación en el Capítulo 10. Allí también el error aleatorio se reducía al aumentar el tamaño de la muestra de acuerdo con $\frac{1}{\sqrt{n}}$.

Los límites de control tres sigma para un gráfico de control $\bar{x}$ son:

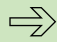
- Límite inferior de control (LIC) = $\mu - 3\,\sigma_{\bar{x}}$
- Límite superior de control (LSC) = $\mu + 3\,\sigma_{\bar{x}}$

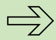
O sea:

- Límite inferior de control (LIC) = $\mu - 3\sigma / \sqrt{n}$
- Límite superior de control (LSC) = $\mu + 3\sigma / \sqrt{n}$

Ejemplo: Continuemos con datos de capacidad volumétrica (sección 21.1.1.1.) para un proceso con media 47 dm^3 y desvío estándar de 0,666 dm^3. Utilizamos esta vez una carta de control $\bar{X}$, con n=5. Es decir, se promedia la capacidad volumétrica de 5 garrafas cada hora, durante 16 horas. Los datos son los siguientes:

Capacidad volumétrica de 5 garrafas						
	Garrafa 1	**Garrafa 2**	**Garrafa 3**	**Garrafa 4**	**Garrafa 5**	**Promedio**
Muestra 1	45,36	46,53	47,36	47,27	46,78	46,66
Muestra 2	47,59	46,10	47,10	47,01	47,52	47,06
Muestra 3	47,44	47,91	46,07	47,11	47,97	47,30
Muestra 4	47,84	46,19	47,01	47,43	46,39	46,97
Muestra 5	46,79	48,21	47,37	46,61	46,39	47,07
Muestra 6	48,11	47,45	46,65	48,01	48,02	47,65
Muestra 7	46,66	47,06	47,95	46,51	46,53	46,94
Muestra 8	46,92	47,87	47,05	47,96	47,18	47,40
Muestra 9	46,59	47,45	45,81	46,55	47,22	46,72
Muestra 10	47,28	46,53	48,17	45,93	47,01	46,98





Capacidad volumétrica de 5 garrafas						
	Garrafa 1	**Garrafa 2**	**Garrafa 3**	**Garrafa 4**	**Garrafa 5**	**Promedio**
Muestra 11	47,18	47,12	47,70	47,09	47,27	47,27
Muestra 12	46,58	47,02	45,82	47,35	46,31	46,62
Muestra 13	46,89	47,39	46,33	47,50	48,18	47,26
Muestra 14	46,17	46,89	46,63	45,00	47,46	46,43
Muestra 15	47,82	47,24	46,86	46,01	47,04	46,99
Muestra 16	46,29	47,59	47,40	45,81	47,62	46,94

Tenemos n=5, **μ**=47 dm³ y **σ**=0,66 dm³, por lo tanto:

- Límite inferior de control (LIC) $= 47 - \dfrac{3 \times 0{,}66}{\sqrt{5}}$
  $= 47-0{,}89$
  $= 46{,}11$

- Límite superior de control (LSC) $= 47 + \dfrac{3 \times 0{,}66}{\sqrt{5}}$
  $= 47-0{,}89$
  $= 47{,}89$

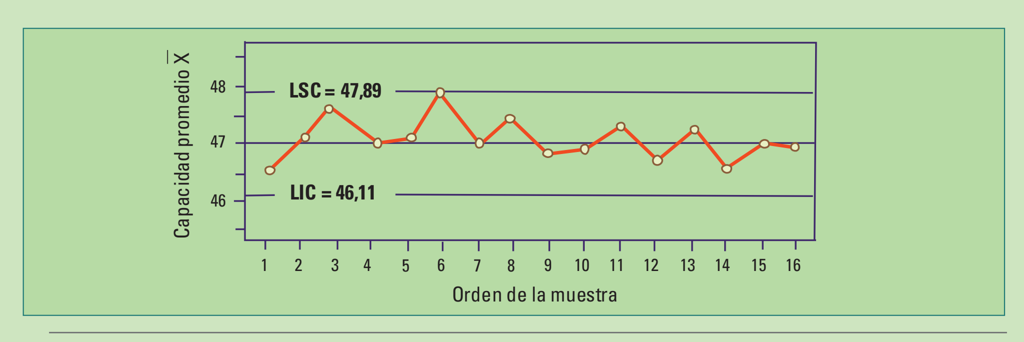


Figura 21.5. *Gráfico de control $\bar{X}$ para la capacidad volumétrica de garrafas, n=5.*

La figura 21.5 no muestra evidencias de que el proceso se haya salido de control, todos sus valores están dentro de las bandas y no aparece ningún patrón no aleatorio.

Nuevo ejemplo: El sector de control de calidad de una fábrica que produce dardos registró los diámetros de 10 submuestras sucesivas de tamaño 4, resultado en los siguientes promedios (en milímetros):

Muestra	1	2	3	4	5	6	7	8	9	10
Promedio	3,01	2,97	3,12	2,99	3,03	3,02	3,1	3,14	3,09	3,2

Registros históricos indican que cuando el proceso opera en control, los diámetros sucesivos tienen distribución gaussiana con media **μ**=3 y desvío **σ**=0.1 por lo tanto para n=4 los límites de control 3-sigma son:

$$LIC = \frac{3 - 3\,(0{,}1)}{\sqrt{4}} \qquad LSC = \frac{3 + 3\,(0{,}1)}{\sqrt{4}}$$

$$LIC = 2{,}85 \qquad LSC = 3{,}15$$

Como la media muestral número 10 se encuentra por encima del límite superior concluimos que hay razones para sospechar que la media de los diámetros de los dardos difiere de 3. Más aún el gráfico de control de la figura 21.6 parece sugerir que a partir de la muestra 6 aumentó la media del diámetro de los dardos.

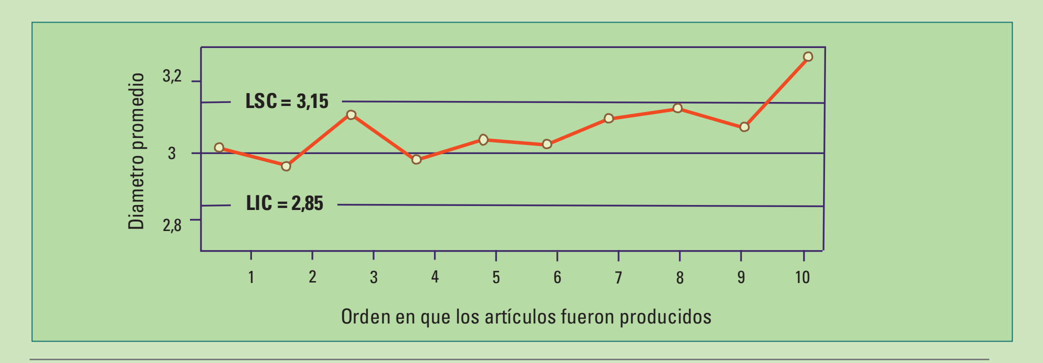


Figura 21.6. *Gráfico de control para diámetros de dardos, de 10 submuestras sucesivas de tamaño 4.*

□ 21.3. Análisis de patrones no aleatorios en cartas de control

Una carta de control puede indicar una condición fuera de control cuando uno ó más puntos caen fuera de los límites de control o cuando los puntos muestran un comportamiento no aleatorio. Por ejemplo, la figura 21.6 muestra un punto fuera de los límites de control pero además desde el punto cuatro hasta el ocho se observa una marcada tendencia creciente con apariencia no aleatoria.

El manual de la empresa Western Electric (Western Electric Handbook (1956)) establece que un proceso está fuera de control cuando se cumple alguna de las siguientes pautas:

1. Hay un punto fuera de los límites 3-**σ**.
2. Dos de tres puntos consecutivos se encuentran del mismo lado fuera de un límite 2-**σ**.
3. Cuatro de 5 puntos consecutivos están se encuentran del mismo lado fuera de un límite 1-**σ**.
4. Ocho puntos consecutivos del mismo lado de la línea central.

Este criterio aumenta la sensibilidad para detectar un proceso fuera de control pero también aumenta la probabilidad de falsa alarma.

La figura 21.7 muestra el gráfico $\bar{X}$ (equis barra) para el ejemplo de los dardos con los límites 1-sigma, 2-sigma y 3-sigma utilizados en el procedimiento Western Electric. A veces estos límites son llamados límites de advertencia. Estos límites dividen el gráfico de control en tres zonas (A, B y C) a cada lado de la línea central. Nótese que los **cuatro últimos puntos** caen en la zona B ó más allá de ella. Luego tenemos en este caso **una doble evidencia de que el proceso no está en control ya que se cumplen las reglas 1 y 3.**

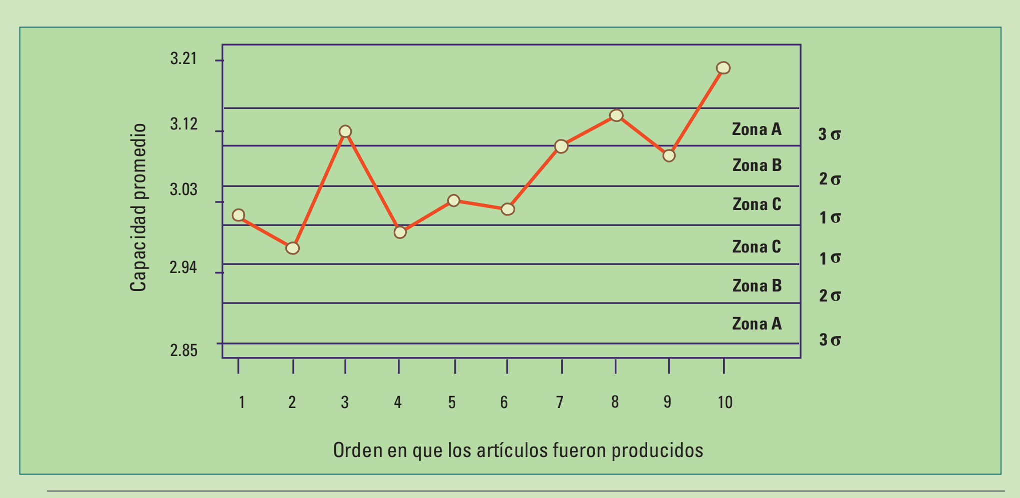


Figura 21.7. *Gráfico de control $\bar{X}$, para diámetros de dardos, con los límites 1-sigma, 2-sigma y 3-sigma.*

22. Relación entre variables

Comencemos con un ejemplo.

Ejemplo 1: En una ciudad con graves problemas de obesidad en la población, se solicitó a un grupo de 60 adolescentes que registrara durante un mes la cantidad de horas que dedicaban cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promediaran. La tabla 22.1 presenta la edad en años (Edad), el género (Varón, Mujer), el promedio de horas por día dedicadas a actividades sedentarias (Horas) y un número (Id) para identificar a cada participante:

PROMEDIO DE HORAS POR DÍA DEDICADAS A MIRAR TELEVISIÓN, ESTUDIAR O UTILIZAR LA COMPUTADORA. TABLA 22.1

Varones						Mujeres					
Id	Edad	Horas	Id	Edad	Horas	Id	Edad	Horas	Id	Edad	Horas
1	11,2	5,5	16	14,6	5,3	31	11,1	4,3	46	13,9	5,0
2	11,4	5,4	17	15,0	5,2	32	11,2	5,1	47	14,2	4,4
3	11,4	4,5	18	15,4	7,0	33	11,2	4,7	48	14,4	5,6
4	11,5	4,8	19	15,6	5,9	34	11,5	4,5	49	14,9	4,4
5	11,6	5,0	20	15,9	6,6	35	11,6	4,7	50	15,1	5,2
6	11,7	5,5	21	16,2	6,3	36	11,6	4,8	51	15,4	5,1
7	11,9	4,3	22	16,5	5,8	37	11,8	4,4	52	15,6	5,1
8	12,6	5,7	23	17,0	6,9	38	11,9	4,7	53	15,9	5,3
9	12,8	4,7	24	17,3	6,9	39	12,3	5,0	54	16,2	4,7
10	13,2	5,4	25	17,4	6,2	40	12,8	4,7	55	16,4	4,9
11	13,8	5,6	26	17,5	5,5	41	12,8	5,1	56	16,6	6,7
12	13,8	5,5	27	17,8	6,0	42	12,9	5,2	57	17,2	5,0
13	14,0	6,6	28	17,9	6,5	43	13,1	5,8	58	17,4	5,8
14	14,3	5,5	29	18,2	6,4	44	13,5	5,2	59	17,9	5,6
15	14,5	5,4	30	18,3	5,7	45	13,6	5,1	60	18,1	5,8

¿Qué información nos pueden brindar los datos de la tabla 22.1? Podemos comparar la distribución de las variables "Edad" y "Horas", en dos grupos definidos por la variable categórica "Género" con dos categorías: Varón, Mujer.

Medidas resumen. TABLA 22.2

	Varones		Mujeres	
	Edad	Horas	Edad	Horas
Mínimo	11,20	4,30	11,10	4,30
1er. Cuartil	12,65	5,33	12,00	4,70
Mediana	14,55	5,65	13,75	5,05
Media	14,68	5,73	14,07	5,06
3er. Cuartil	16,88	6,38	15,82	5,20
Máximo	18,30	7,00	18,10	6,70

No se observan diferencias llamativas entre las distribuciones de las edades de varones y mujeres, tanto mirando la tabla 22.2 como la figura 22.1. Sin embargo para las horas, todos los valores para los varones entre el 1er cuartil y el máximo (5,33-7,00) son mayores que todos los valores para las mujeres entre el mínimo y el 3er. cuartil (4,30-5,20). Esto sugiere que en general los adolescentes de esta ciudad le dedican más horas a las actividades sedentarias que las adolescentes.

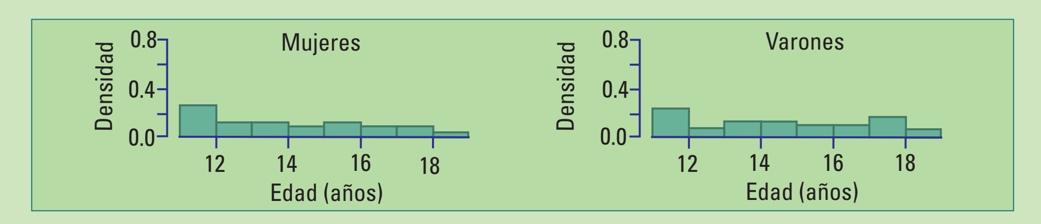


Figura 22.1. *Histogramas de las edades de mujeres y varones.*

Para las mujeres la media muestral de la cantidad de horas por día dedicadas a actividades sedentarias es 5,06. En los varones es 5,73 horas, aproximadamente 3/4 de hora más. La figura 22.2 refuerza esta situación. Los intervalos correspondientes a la mayor cantidad de horas (de 6 a 7) presentan mayor densidad de datos para los varones que para las mujeres. El intervalo más poblado para las mujeres es entre 4,5 y 5,0 horas y en los varones entre 5 y 5,5 horas.

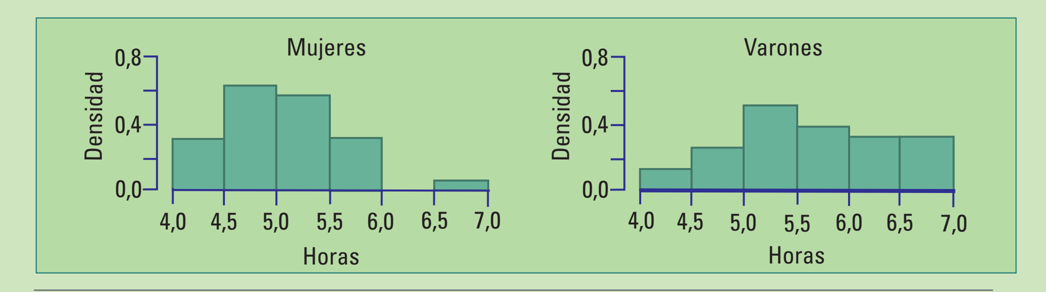


Figura 22.2. *Histogramas de las horas por día dedicadas a actividades sedentarias de mujeres y varones.*

El gráfico caja de la figura 22.3 para mujeres muestra además un valor atípico; se trata de un valor muy alejado del resto. Se destaca también que la caja correspondiente a los varones se encuentra desplazada hacia arriba (hacia los valores mayores) en comparación con las mujeres; por lo tanto, más del 75% de los valores de horas para los varones son mayores que más del 75% de los valores menores de las horas para mujeres. Esto es lo mismo que notamos al describir las medidas resumen.

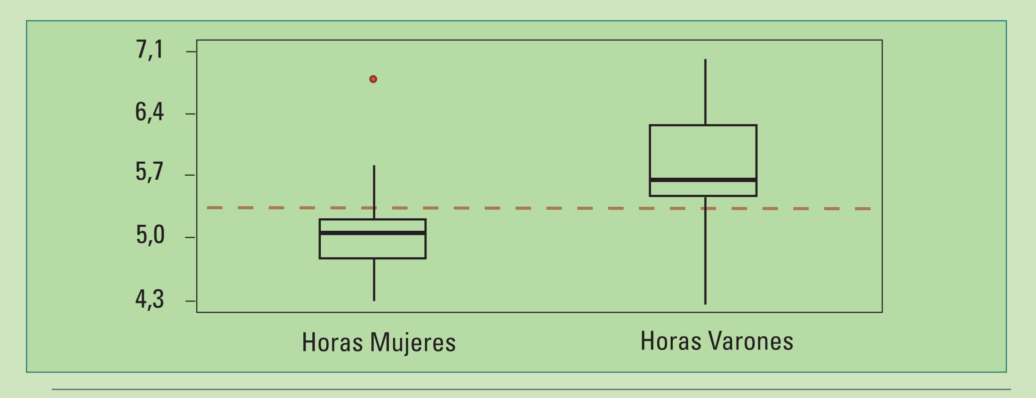


Figura 22.3. *Gráficos caja para la cantidad de horas por día dedicadas a actividades sedentarias de varones y mujeres. Se destacan un valor atípico y los valores de las cajas que no se superponen. La caja para mujeres se encuentra por debajo de la de los varones.*

Hasta aquí comparamos los valores de una variable continua por vez en dos grupos definidos por una variable categórica.

Nos preguntamos ahora:

- ¿Habrá alguna relación entre la cantidad de horas dedicadas a actividades sedentarias y la edad?
- ¿Los más chicos le dedicarán mayor o menor cantidad de horas que los más grandes a ese tipo de actividades?

Se trata en este caso de relacionar los valores de dos variables cuantitativas continuas (horas, edad).

□ 22.1. Diagrama de dispersión

La forma gráfica más habitual de describir la relación entre dos variables cuantitativas es utilizando un **diagrama de dispersión.** Cada **punto** corresponde a **un par de valores** (uno para cada variable), medidos sobre el mismo individuo.

En general, si una de las variables puede pensarse como explicativa de la otra (**variable explicativa**), siempre se la grafica en el eje horizontal (eje x) y la otra (**variable respuesta**) en el eje vertical (eje y).

En el ejemplo de la figura 22.4, la edad es la **variable explicativa.** Pensamos que la edad puede explicar, aunque sea en parte, la cantidad de horas diarias dedicadas a actividades sedentarias (**variable respuesta,** graficada en el eje y).

Cada punto representa a un varón. Está determinado por su edad y la cantidad de horas diarias dedicadas a las actividades sedentarias relevadas (mirar televisión, estudiar o utilizar la computadora).

Para ilustrar como se realiza el gráfico, se destaca el punto correspondiente a Id = 13, Edad =14, Horas = 6,6.

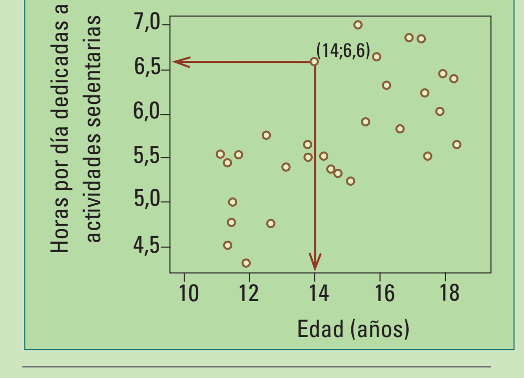


Figura 22.4. Diagrama de dispersión de las horas en función de la edad para los varones. Se destaca el punto correspondiente a Id = 13, Edad =14, Horas = 6,6.

En un diagrama de dispersión observamos el patrón general de la relación entre las variables mirándolo de izquierda a derecha.

Si a medida que **x aumenta** (es decir, nos corremos hacia la derecha del gráfico):

- en promedio también lo hace y (los valores de y se encuentran más arriba); esto indica una **asociación** lineal **positiva** entre las variables.
- en promedio y decrece (los valores de y se encuentran más abajo, esto indica una **asociación** lineal **negativa** entre las variables.
- no puede determinarse una tendencia de crecimiento o decrecimiento en los valores de y; esto significa que no hay una asociación lineal entre las variables.

Más formalmente diremos:

- Dos variables están asociadas en forma **positiva** cuando los valores que están por **encima del promedio** de una de ellas tienden a ocurrir mayoritariamente con valores por **encima del promedio** de la otra. Lo mismo ocurre con los que se encuentran por debajo del promedio.
- Dos variables están asociadas en forma **negativa** cuando valores por **encima del promedio** de una suelen estar acompañados por valores por **debajo del promedio** de la otra y viceversa.

La figura 22.5 muestra el diagrama de dispersión dividido en cuatro cuadrantes determinados por la media muestral de las horas (= 5,73) y la media muestral de las edades (= 14,68).

¿Cuántos puntos hay en cada uno de los cuadrantes de la figura 22.5? 11 puntos en el Cuadrante I, 2 puntos en el Cuadrante II, 15 puntos en el Cuadrante III y 3 puntos en el Cuadrante IV.

La mayoría de los puntos se encuentran en el primer y tercer cuadrante. **En el primero** las edades están **por encima de su media muestral** y lo mismo ocurre con las horas. **En el tercero** tanto las edades como las horas son **menores** que **sus respectivas medias** muestrales. Por lo tanto se trata de una **asociación** lineal **positiva.**

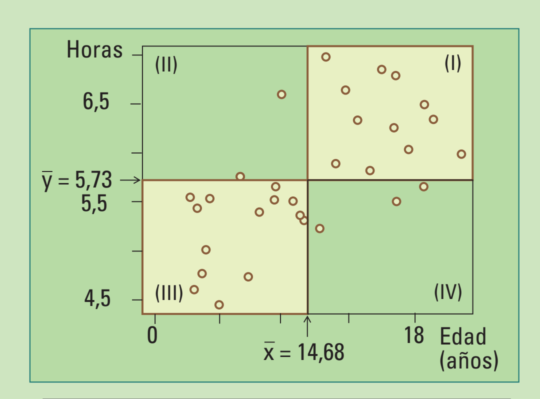


__Figura 22.5.__ Diagrama de dispersión de las horas en función de la edad para los varones y los cuatro cuadrantes determinados por las medias muestrales de las edades y las horas.

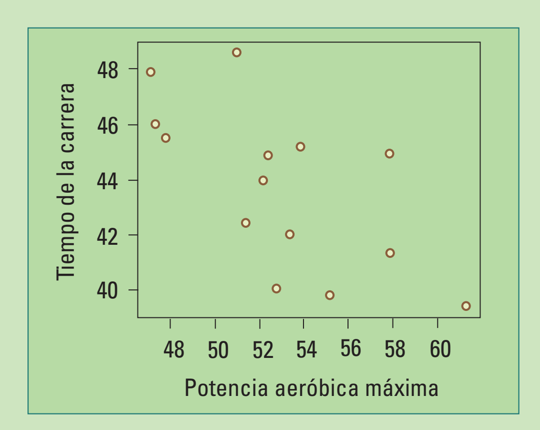


__Figura 22.6.__ Diagrama de dispersión de Tiempo (Y, min) en 10 km y la Potencia aeróbica máxima (X; ml kg^{-1} min^{-1}) de 14 atletas entrenadas. A medida que aumenta la potencia aeróbica máxima alcanzada disminuye el tiempo.

Ejemplo 2: Los datos de la tabla 22.3 (Atletas) corresponden a un estudio sobre la relación entre el grado de entrenamiento y el desempeño posterior en una carrera de 10 km. Se evaluaron 14 mujeres entrenadas. El grado de entrenamiento se mide mediante la "Potencia aeróbica máxima" (ml/(kg min) alcanzada y el desempeño posterior mediante el tiempo empleado en completar 10 km durante una competencia.

Tiempo (Y, min) en 10 km y la potencia aeróbica máxima (X, ml kg^{-1} min^{-1}). TABLA 22.3

Atleta	x	y
1	47,170	47,830
2	47,410	46,030
3	47,880	45,600
4	51,050	48,550
5	51,320	42,370
6	52,180	43,930
7	52,370	44,900
8	52,830	40,030
9	53,310	42,030
10	53,930	45,120
11	55,290	39,800
12	57,910	44,900
13	57,940	41,320
14	61,320	39,370

A medida que **aumenta la Potencia** aeróbica máxima (X) alcanzada por las atletas, durante el entrenamiento previo a la competencia, mejora su rendimiento en la carrera de 10 km, **disminuyendo el tiempo** (Y) (figura 22.6). Por lo tanto, X e Y están asociadas negativamente.

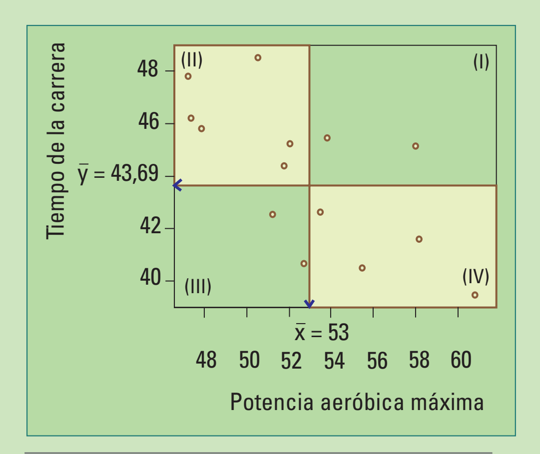


__Figura 22.7.__ Tiempo (Y, min) en 10 km y la Potencia aeróbica máxima (X; ml kg^{-1} min^{-1}) de 14 atletas entrenadas. Los cuadrantes II y IV contienen 10 de los 14 puntos del diagrama de dispersión, mostrando la asociación negativa entre X e Y.

¿Cuántos puntos hay en cada uno de los cuadrantes de la figura 22.7?
2 puntos en el Cuadrante I, 6 puntos en el Cuadrante II, 2 puntos en el Cuadrante III y 4 puntos en el Cuadrante IV.

Esta vez, los cuadrantes II y IV, determinados por las medias muestrales de cada una de las variables, con 10 de los 14, contienen la mayoría de los puntos. Reafirmamos que el tiempo en realizar la carrera de 10 km y el nivel de entrenamiento medido por la potencia aeróbica máxima tienen **asociación negativa.**

En general, un diagrama de dispersión muestra la forma, la dirección y el grado de la asociación entre los valores de dos variables cuantitativas. Cuanto más cerca se encuentren los puntos del diagrama de una recta tanto más fuerte es la asociación lineal entre los valores de las variables graficadas. Pero nuestra percepción visual del grado de asociación puede estar equivocada debido a la escala. La figura 22.8. muestra el diagrama de dispersión del mismo conjunto de datos (Atletas) en diferentes escalas. En el diagrama de la izquierda la asociación lineal parece más fuerte en comparación con el de la derecha.

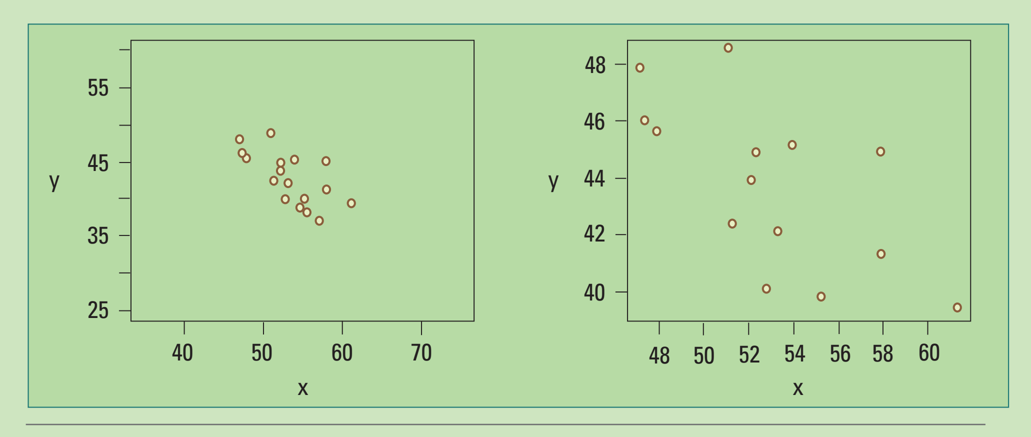


__Figura 22.8.__ Dos diagramas de dispersión de los mismos datos (tabla 22.3). El de la izquierda sugiere una asociación más fuerte entre las variables que el de la derecha.

22.2. Coeficiente de correlación

Necesitamos un número que no dependa de las escalas del gráfico de dispersión y represente el grado de asociación lineal entre los pares de valores de dos variables continuas.

¿Qué propiedades debería tener ese número?

- Ser positivo si la asociación lineal es positiva.
- Ser negativo si la asociación lineal es negativa.
- Ser más grande, en valor absoluto, cuánto más cerca se encuentren de alguna recta los pares de valores.
- No depender de las unidades en las que se expresan las variables.

Un número con todas las propiedades anteriores es el Coeficiente de Correlación de Pearson (r):

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

donde $(x_1, y_1)...(x_n, y_n)$, es un conjunto de **pares de datos** de tamaño **n**, correspondiente a observaciones de **dos variables** continuas **X** e **Y.**

Como otras de las fórmulas de cálculo de estadística, **¡asusta!,** pero no es problema para las calculadoras y menos aún para las computadoras.



¿Coeficiente de correlación de Pearson?
Su fórmula me recuerda a la del desvío estándar muestral.

Otra forma de escribir elcoeficiente de correlación es: $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

donde

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \quad \textbf{y} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

son los **desvíos estándar muestrales de X e Y respectivamente.**

¿Cuánto vale r en el ejemplo 1, de las atletas?

r = -0,659

Como esperábamos es negativo. Su valor absoluto (0,659) es menor a 1.

Más propiedades del coeficiente de correlación muestral r:

- No **depende de las unidades** en que se miden las variables y su valor está siempre entre -1 y 1.
- No distingue entre variable explicativa (**X**) y variable respuesta (**Y**): el coeficiente de correlación entre **X** e **Y** es igual al coeficiente de correlación entre **Y** y **X.**
- A mayor valor absoluto de r, mayor el grado de **asociación lineal.**
- Cuando r = 0 no hay una tendencia lineal creciente o decreciente en la relación entre los valores x´s e y´s.
- Los valores extremos, r = 1 y r = -1, ocurren únicamente cuando **los puntos en un diagrama de dispersión caen exactamente sobre una recta.** Corresponde a asociaciones positivas ó negativas perfectas.
- Valores de **r cercanos a 1** ó **-1** indican que los puntos yacen **cerca** de una recta.

Como el **denominador de r es siempre positivo** $\left(\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)$, para comprender de donde se obtiene su signo, sólo es necesario estudiar el signo del numerador $\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)$

- Cuando la mayoría de los sumandos son positivos: $(x_i-\bar{x})(y_i-\bar{y}) > 0$ la suma

  $$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

  es positiva y por lo tanto **r es positivo.** Ocurre cuando la mayoría de los puntos (x_i, y_i) se encuentran en los cuadrantes (I) y (III). En esos cuadrantes los desvíos $\mathbf{x_i-\bar{x}}$ e $\mathbf{y_i-\bar{y}}$ tienen **el mismo signo** y su producto es positivo.
- Cuando la mayoría de los sumandos son negativos: $(x_i-\bar{x})(y_i-\bar{y}) < 0$, o sea cuando los puntos (x_i, y_i) se encuentran en su mayoría en los cuadrantes (II) y (IV), allí los desvíos $\mathbf{x_i-\bar{x}}$ e $\mathbf{y_i-\bar{y}}$ tienen **signos opuestos** y su producto es negativo. La suma $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ resulta negativa y por lo tanto **r es negativo.**

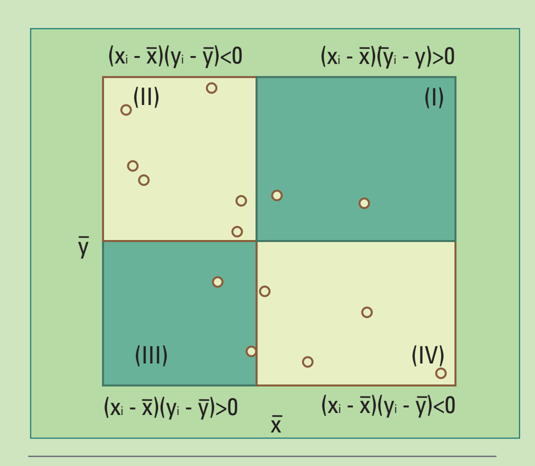


Figura 22.9. *El signo del producto $(x_i - x)(y_i - y)$ es positivo para los puntos de los cuadrantes (I) y (III); negativos en los otros dos cuadrantes.*

Cuantos más puntos caigan en los cuadrantes (I) y (III) habrá **más sumandos** de

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

positivos contribuyendo al total positivo **r > 0**

Cuantos más puntos caigan en los cuadrantes (II) y (IV) habrá **más sumandos** de

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

negativos contribuyendo al total positivo **r < 0**

La figura 22.10 muestra cómo los valores de r decrecen en valor absoluto, se alejan del 1 ó -1, a medida que decrece el grado de asociación lineal entre las variables.

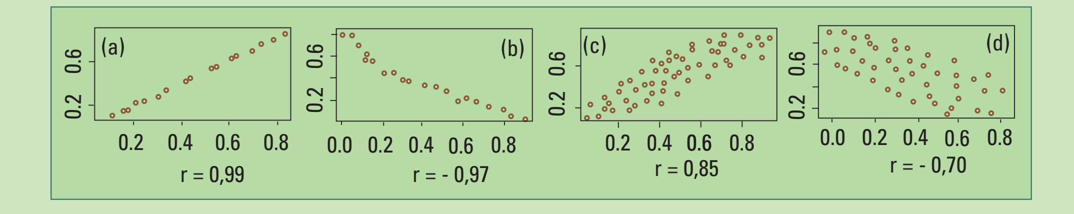
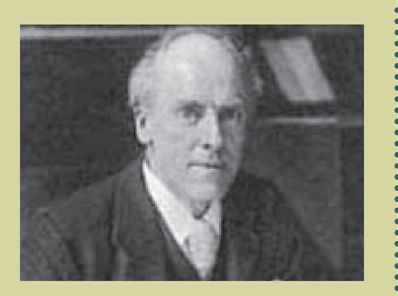


Figura 22.10. *En (a) y (b) se muestran datos con un alto grado de asociación lineal, en el primero la asociación es positiva y en el segundo negativa. En (c) y (d) los datos están menos concentrados sobre una recta pero sigue habiendo claras tendencias: creciente (c) y decreciente (d).*

Karl Pearson (1837 -1936) Estadístico, historiador y pensador británico. Realizó una intensa investigación sobre desarrollo y aplicación de métodos estadísticos a problemas provenientes de la biología. En 1911 fundó el primer departamento de estadística en la Universidad de Londres.



El coeficiente de correlación de Pearson r mide el **grado de asociación lineal** entre los pares de valores de dos variables continuas. Es un número entre -1 y 1. Su signo refleja si la asociación es positiva o negativa y su valor está más acerca de 1 (ó -1) a medida que los puntos del diagrama se aproximan a una recta. Los valores extremos 1 y -1 se obtienen cuando los puntos del diagrama de dispersión están perfectamente alineados.

A veces el coeficiente de correlación no refleja lo esperado, como vemos en (b) y (c) de la figura 22.11. Ante una relación curvilínea (b) el coeficiente de correlación (r = -0.01) indica que no hay asociación entre las variables a pesar de existir una **asociación no lineal** muy fuerte. En (c) el coeficiente de correlación (r = -0.75) indicaría una asociación negativa cuando en realidad se trata de dos grupos de datos, uno con asociación positiva y el otro con asociación nula.

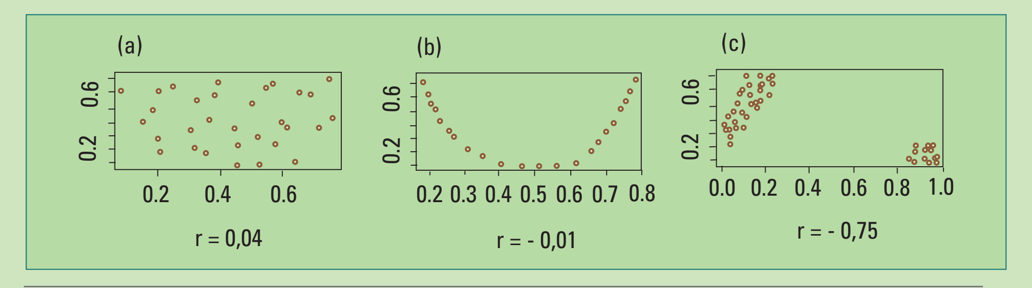


Figura 22.11. *El coeficiente de correlación es cercano a cero cuando los pares de datos no están asociados (a), pero también puede ser nulo o casi nulo ante una relación no lineal entre los valores de X e Y (b).*

□ 22.3. Recta de regresión lineal simple

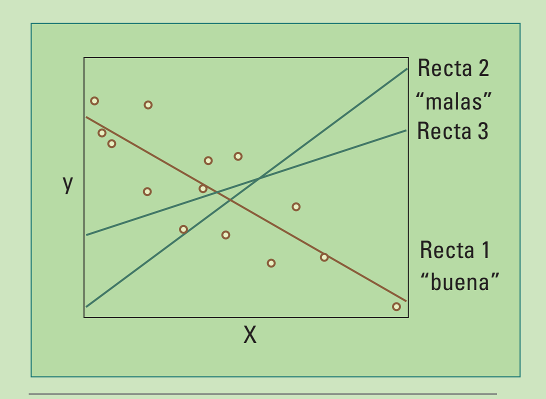


Fig. 22.12. *Tres rectas en un diagrama de dispersión.*

Cuando un diagrama de dispersión muestra un patrón lineal es deseable resumir ese patrón mediante la ecuación de una recta. Esa recta debe representar a la mayoría de los puntos del diagrama, aunque ningún punto esté sobre ella.

La recta 1 de la figura 2.12. representa bien la dirección y el sentido de la asociación entre los valores de **X** e **Y**, pasa "cerca" de la mayoría de los puntos del diagrama de dispersión; decimos que es una recta "buena". Esto no ocurre con las rectas 2 y 3. Pero, ¿cómo podemos elegir la mejor de las rectas posibles?

22.3.1. Cuadrados mínimos

El **método de cuadrados mínimos** propone elegir la recta que **minimiza** la suma de los cuadrados de las **distancias verticales** de cada punto a la recta.

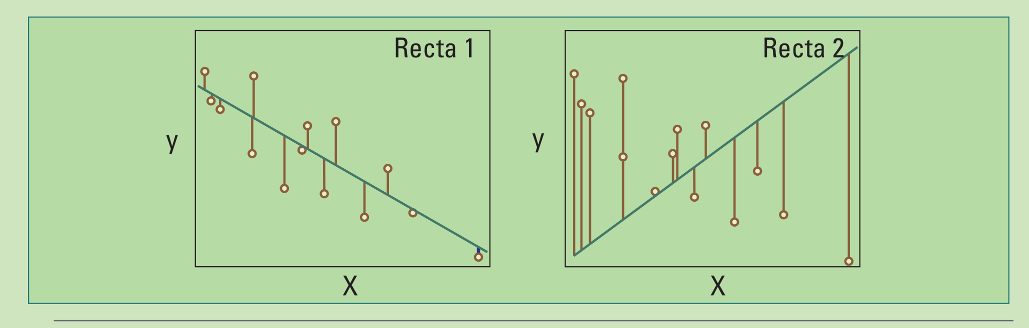


Figura 22.13. *Dos diagramas de dispersión del mismo conjunto de pares de datos y las distancias verticales a dos rectas.*

Los diagramas de dispersión de la figura 22.13 muestran **los mismos puntos.** Las dis**tancias verticales** de los puntos a la recta 1 son, en su mayoría, menores que a la recta 2. Por lo tanto la suma de los cuadrados de esas distancias será menor para la recta 1 que para la recta 2.

Consideremos la ecuación de una recta cualquiera **y=a+bx.** Sean:

- (x_i, y_i) las coordenadas de un punto del plano representando al dato i;
- $(x_i, \hat{y}_i)$ las coordenadas de un **punto sobre la recta** con $x = x_i$

donde ŷ (se lee y sombrero sub i) se obtiene reemplazando x_i en la ecuación de la recta ($\hat{y}_i = a+bx_i$, figura 22.14).

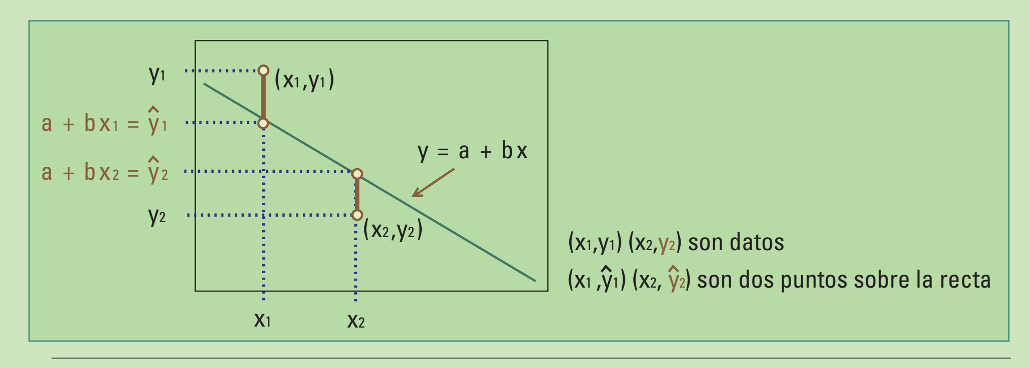


Fig. 22.14. *Dos puntos en el plano y una recta. El primero está por encima de la recta, tiene residuo positivo* ($y_1 - \hat{y}_1 > 0$)*. El segundo está por debajo de la recta, tiene residuo negativo* ($y_2 - \hat{y}_2 < 0$)*.*

La distancia vertical de un punto (x_i, y_i) a la recta es llamada residuo y se obtiene de la siguiente manera:

$$\text{residuo}_i = y_i - \hat{y}_i$$
$$= y_i - (a+bx_i)$$

Algunos **residuos** son **positivos,** la respuesta observada está por encima de la recta, y otros son **negativos,** la respuesta observada está por debajo de la recta. La figura 22.15 muestra esta situación:

- El primer punto se encuentra por encima de la recta $(y_1 > \hat{y}_1)$ luego $y_1 - \hat{y}_1 > 0$.
- El segundo punto se encuentra por debajo de la recta $(y_2 > \hat{y}_2)$ luego $y_2 - \hat{y}_2 < 0$.

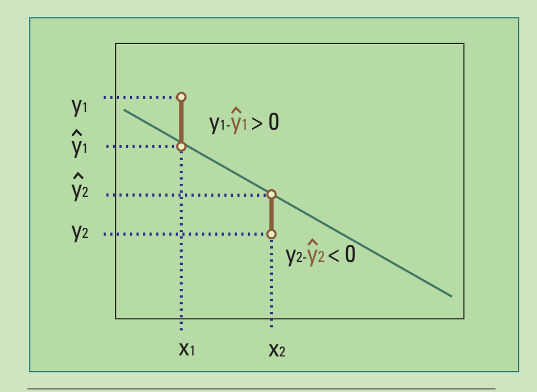


Fig. 22.15. *Dos residuos uno positivo y otro negativo.*

Pero ¿cómo hallamos los coeficientes ***a y b*** de la recta que minimiza la suma de los cuadrados de los residuos?

Mediante el **método de cuadrados mínimos** (CM) ***a y b*** se eligen de manera que **la suma de los cuadrados de los residuos**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (a + b x_i) \right)^2$$

sea mínima.

Los coeficientes de la recta estimada por CM se calculan, a partir de los datos, mediante las siguientes ecuaciones:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

La deducción está fuera del alcance de este texto, sin embargo es interesante lo que estas ecuaciones nos muestran:

- La primera ecuación dice $b = r \frac{s_y}{s_x}$, donde **r** es el coeficiente de correlación y $\mathbf{s_x}$ y $\mathbf{s_y}$ son los desvíos estándar muestrales de las x's y las y's respectivamente. Por lo tanto si $s_x = s_x = 1$ resulta que la pendiente de la recta ajustada es igual al coeficiente de correlación.
- La segunda ecuación nos dice que **la recta de cuadrados mínimos pasa por** el punto ($\bar{x}$, □) pues sus coordenadas satisfacen la ecuación de la recta ajustada (□ = $a + \bar{b}x$).

- La suma de los residuos es 0 y equivalentemente su promedio: Si sumamos los residuos, y_1 - $(a+bx_1)$, y los dividimos por n entonces,

$$\frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} (a+bx_i)}{n} = \bar{y} - (a+b\bar{x}) \qquad \text{pues } \mathbf{a = \hat{y}-b\bar{x}}$$

$$= 0$$

Ninguna otra recta tendrá, para el mismo conjunto de datos, una suma de cuadrados de los residuos tan baja como la obtenida por CM. En este sentido, el método de mínimos cuadrados brinda la solución que mejor ajusta a un conjunto de datos.

En la práctica, la recta de CM se obtiene utilizando una calculadora o, mejor aún, una computadora.

Ejemplo 3: Retomemos los datos de la tabla 22.3 (Atletas) considerando al Tiempo (Y) como variable respuesta y a la Potencia aeróbica máxima (X) como variable explicativa. Pensamos que el grado de entrenamiento medido por la variable X puede explicar, aunque sea parcialmente, el tiempo realizado en una carrera de 10 km.

Calcularemos la pendiente b y la ordenada al origen a mediante las fórmulas:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Obtenemos los siguientes resultados utilizando una calculadora o una computadora: $\bar{x}=52{,}99 \quad \bar{y}=43{,}70$

$$\sum_{i=1}^{14} (x_i - \bar{x})(y_i - \bar{y}) = -104{,}39 \qquad \sum_{i=1}^{14} (x_i - \bar{x})^2 = 223{,}11$$

Por lo tanto

$$b = \frac{\sum_{i=1} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{-104{,}39}{223{,}11}$$

$$= -0{,}468$$

$$a = 43{,}70 - (-0{,}4678)\ (52{,}99) = 68{,}49$$

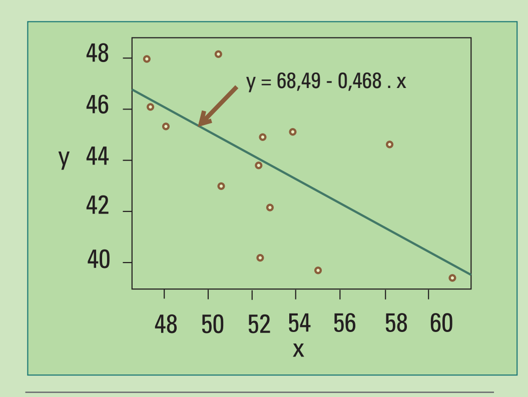


Fig. 22.16. *Diagrama de dispersión de los datos Atletas y la recta de regresión simple ajustada por el método de cuadrados mínimos.*

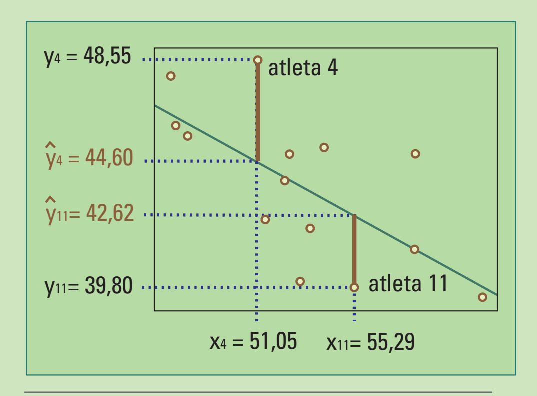


Fig. 22.17. *El punto correspondiente a la atleta 4 se encuentra por encima de la recta ajustada por cuadrados mínimos (su residuo es positivo). El de la atleta 11 por debajo, su residuo es negativo .*

La ecuación de la recta ajustada a los datos de las atletas (figura 22.19) es:

$$y = 68{,}49 - 0{,}468\ x$$

Algunos residuos son positivos y otros negativos (figura 22.17) :

Para la atleta 4, el punto en el diagrama de dispersión se encuentra **por encima de la recta.** El **residuo** es **positivo,**

$$y_4 - \hat{y}_4 = 48{,}55 - 44{,}60$$
$$= 3{,}95 > 0$$

$$\hat{y}_4 = 48{,}55$$
$$\hat{y}_4 = 68{,}49 - 0{,}468 \times 51{,}05$$
$$= 44{,}60$$

Para la atleta 11, el punto en el diagrama de dispersión se encuentra **por debajo de la recta.** El **residuo** es **negativo** y

$$y_{11} - \hat{y}_{11} = 39{,}80 - 42{,}62$$
$$= -2{,}82 < 0$$

$$\hat{y}_{11} = 39{,}80$$
$$\hat{y}_{11} = 68{,}49 - 0{,}468 \times 55{,}29$$
$$= 42{,}62$$

En las secciones siguientes veremos como la recta ajustada en un diagrama de dispersión de dos variables X e Y resume la relación entre las mismas

22.3.2. Significado de la recta

Veamos qué significa que dos variables **X** e **Y** tengan una relación "perfectamente lineal". Esto es, los pares de (x, y) de los valores de las variables **siempre** se encuentren sobre una recta.

La ecuación de una recta con pendiente b y ordenada al origen a es: y = a + b x. Cuando "x" aumenta una unidad "y" crece (o decrece) b unidades; a es el valor donde la recta corta al eje vertical (figura 22.18).

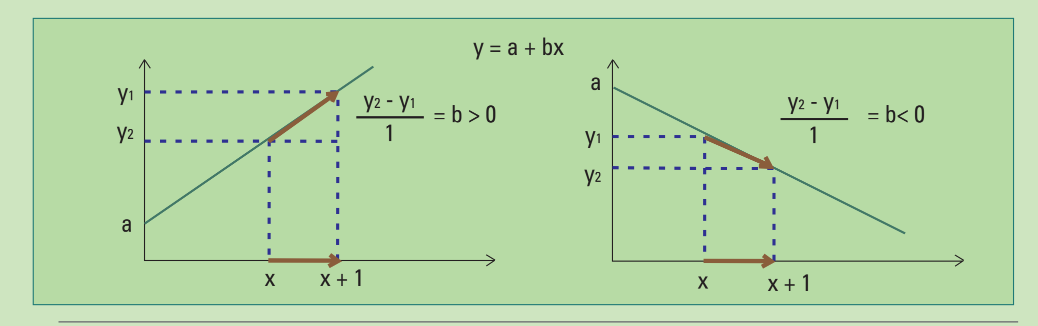


Fig. 22.18. *Dos rectas. La de la izquierda tiene pendiente positiva y la de la derecha pendiente negativa.*

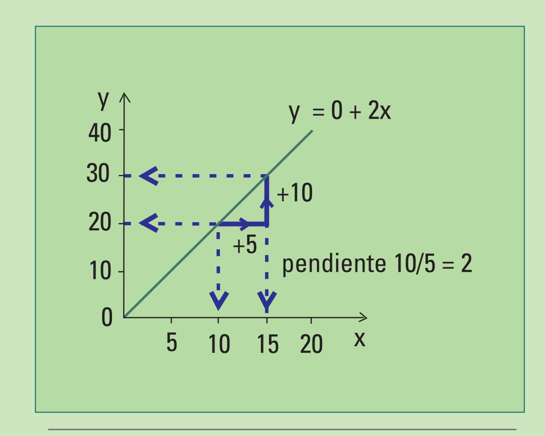


Figura 22.19. *Relación, hipotética, lineal entre la variable "vocabulario en miles de palabras" (y) y **la variable edad** (x).*

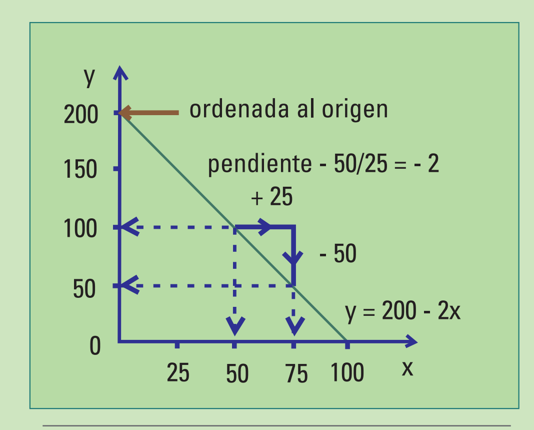


Figura 22.20. *Relación lineal entre la cantidad de palabras recordadas en una prueba **(y)** y la edad **(x)**.*

Ejemplo 4: Supongamos que la relación entre la variable "vocabulario en miles de palabras" (**Y, variable respuesta**) y la **variable edad** (X, variable **explicativa**) satisface la ecuación de la recta con **ordenada al origen a** = 0 y **pendiente b** = 2:

Esta recta corta al eje **Y** en el punto (0,0) y aumenta 2 unidades verticalmente (**Y**) con cada aumento horizontal (**X**) de 1 unidad. En términos de la relación entre la variable edad y el vocabulario esto significa que 0 es la cantidad inicial de palabras (a los 0 años) y que hay un **aumento de 2000** palabras **por año,** aproximadamente (7 por día). La pendiente indica la tasa de cambio. Como **la pendiente es positiva, la relación** entre el vocabulario y la edad **es creciente** (figura 22.19).

Ejemplo 5: En la figura 22.20 **y** representa la cantidad de palabras recordadas en una prueba y nuevamente **x** es la edad. La recta tiene ordenada al origen a = 200 y pendiente ***b* = *-2:***

$y = 200 - 2x$;

corta el eje Y en el punto (0,200) y cae 2 unidades verticalmente con cada aumento de 1 unidad en X. A medida que **aumenta la edad** en un año la **cantidad de palabras** que se recuerda en la prueba **disminuye** en 2.

Como **la pendiente es negativa,** la relación entre la cantidad de palabras recordadas y la edad **es decreciente.** La ordenada al origen, de 200 palabras para edad = 0, no tiene ningún significado biológico como ocurre muchas veces.

En general, **si la relación entre X e Y es perfectamente lineal** y conocemos los valores a y b, la ecuación y = a + b x permite predecir qué valor de Y corresponde a cualquier valor de X. Se trata de una **asociación determinística.** Más aún, dos pares de datos son suficientes para hallar los parámetros a y b, de la misma manera que **dos puntos y una regla alcanzan para dibujar una línea recta.**

22.3.3. Modelo

La relación entre datos reales rara vez es tan simple como la expresada por la ecuación de la recta.

Un modelo más realista plantea que **la media poblacional** de **Y**, más que los valores individuales, **cambia linealmente con X**

$$\mu_Y(x) = \alpha + \beta x$$

En el caso las atletas (tabla 22.3), el modelo de regresión lineal dice que para **cada valor** de **la potencia aeróbica máxima** (**x**), **la media de los tiempos** (**Y**) en una carrera de 10 km, en la población de todas las atletas entrenadas, es:

Tiempo medio (depende de x) = $\boldsymbol{\alpha}$ + $\boldsymbol{\beta}$. (Potencia aeróbica máxima)

Otras cosas, además de X, hacen que los valores individuales Y varíen alrededor de la media (**μ_Y(x)**) de todos los valores de Y cuando X toma el valor **x.** Por ejemplo, además del grado de entrenamiento (medido por la Potencia aeróbica máxima), **otros factores** causan que los **tiempos individuales varíen** alrededor de su media poblacional: **la edad, la longitud de piernas y brazos, la fuerza de piernas y brazos, la motivación por ganar,** etc.

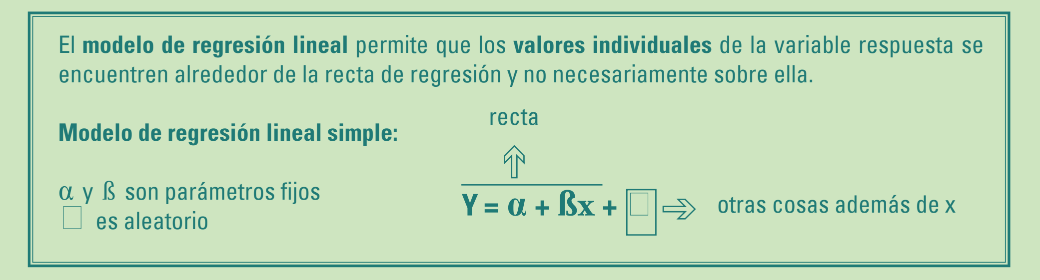
Representamos esas "otras cosas" con **un término de error,** ε (epsilon). Es la diferencia entre un valor individual y la media de la variable **Y** en la población, para un valor fijo **x** de la variable explicativa:

$$\varepsilon = Y - \mu_Y(x)$$
$$\varepsilon = Y - (\alpha + \beta x)$$

Despejando Y de la primera y de la segunda de las igualdades anteriores resultan: $Y = \mu_Y(x) + \varepsilon$

y equivalentemente $Y = \alpha + \beta x + \varepsilon$

El valor de Y es igual a la media más un error, distinto para cada valor de la variable.



En la sección 15.5 vimos el siguiente **modelo de medición** (sin sesgo):

$$Y = \mu_Y + \text{error aleatorio},$$

donde Y representa un valor cualquiera de una variable definida en una población y μ es la media de todos los valores posibles de esa variable en la población considerada. Por ejemplo Y = tiempo que tarda un atleta cualquiera en una carrera de 10 km, μ_Y = media de los tiempos que tardan todas las atletas de una población. En este modelo no se tiene en cuenta la edad y se establece que el **tiempo de una atleta** es igual a la **media de los tiempos de toda la población** más un término llamado de **error.**

Observemos que el **modelo de regresión lineal:**

$$Y = \mu_Y(x) + □$$

tiene una estructura similar al modelo de medición, pero ahora **la media poblacional depende linealmente de una variable explicativa** (x).

El término de error, en cualquiera de los dos modelos permite describir la variabilidad de las observaciones alrededor de la media poblacional (μ_Y o $\mu_Y(x)$). Muchas veces se utilizan curvas Normales para describir esa variabilidad.

Figura 22.21. *Modelo de regresión lineal simple. Para cada valor de x, los valores de la variable Y se distribuyen alrededor de la media $\mu_Y(x) = \alpha+\beta x$.*

Veamos qué ocurre con el modelo de regresión lineal:

$$y = \mu_y(x) + □$$

Utilizamos la curva Normal para describir la variabilidad de las observaciones alrededor de la media (figura 22.21) y tenemos una media diferente para cada valor de la variable explicativa. Sobre la recta de regresión, el valor de y esta determinado por x; y = α + ßx.

Se trata de medias poblacionales, pero no conocemos ni α ni ß.

La recta de regresión verdadera, $\mu_Y(x) = \alpha + ßx$, vincula la media poblacional de la variable respuesta con la variable explicativa. La ordenada al origen (α) y la pendiente (ß) de esta recta son desconocidos.

22.3.3.1 Interpretación de los coeficientes estimados

Los coeficientes a y b obtenidos por el método de cuadrados mínimos (sección 22.3.1) estiman los parámetros α y ß.

Ejemplo 6: Sigamos con los datos de la tabla 22.3 (Atletas) considerando al Tiempo (**Y**) como variable respuesta y a la Potencia aeróbica máxima (**X**) como variable explicativa. Vimos (ejemplo 3) que la ecuación de la recta ajustada a los datos de las atletas (figura 22.16) resulta y = 68,49 - 0,468 x.

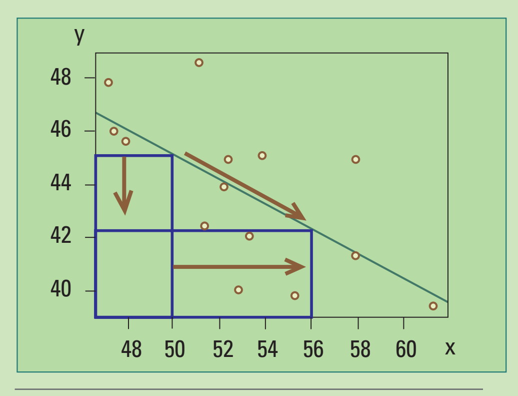
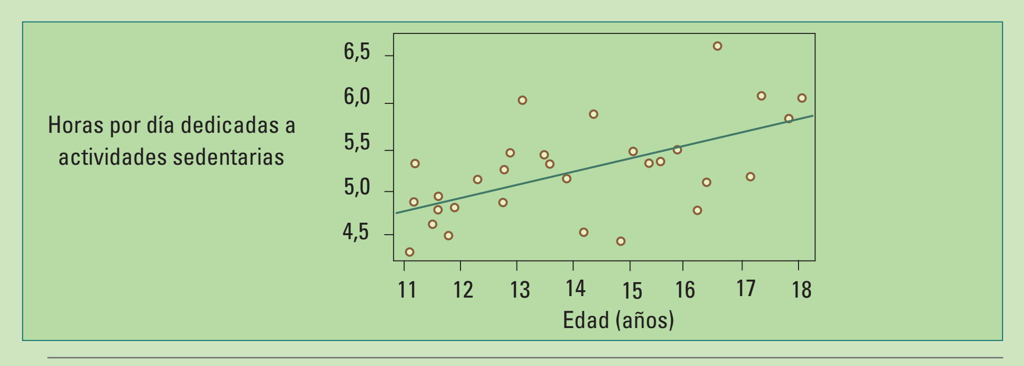


Fig. 22.22. Sobre la recta de regresión ajustada a los datos "Atleta", un aumento de 6 unidades en x resulta en un reducción de 2,81 unidades en y.

La pendiente ajustada es b = -0,468, esto significa **que sobre la recta ajustada,** un aumento de una unidad en x produce un descenso 0,468 unidades en y. Estimamos una reducción de la **media del tiempo** en una carrera de 10 km en 0,468 minutos, cada vez que la **Potencia aeróbica máxima aumenta una unidad.** Esta reducción del tiempo podría no tener importancia. Quizás sea más interesante la reducción del tiempo cuando la Potencia aeróbica máxima aumenta 6 unidades, esto es una **reducción** de 6 . 0,468 = 2,81 minutos en el tiempo de la carrera (figura 22.22).

! a=68,49 y b=-0,468 son valores de estadísticos, es decir, estimaciones de los parámetros poblacionales α y ß.



***22.23.** Diagrama de dispersión de las horas dedicadas a mirar televisión, estudiar o utilizar la computadora en función de la edad para 30 mujeres adolescentes, junto con la recta de regresión ajustada por cuadrados mínimos.*

Ejemplo 7: Consideremos nuevamente los datos de la tabla 22.1 horas por día dedicadas a mirar televisión, estudiar o utilizar la computadora, esta vez para las mujeres.

Tomando como variable respuesta Y = "horas por día dedicadas a mirar televisión, estudiar o utilizar la computadora" y variable explicativa **X** = edad, obtenemos la siguiente recta ajustada por cuadrados mínimos a los datos de la tabla 22.1 (mujeres):

$$Y = 3{,}24 + 0{,}13.\text{edad}$$

El signo de la pendiente de la recta ajustada **es positiva,** mostrando una asociación positiva entre las horas y la edad.

El valor de la pendiente es el **cambio de y cuando la variable explicativa aumenta una unidad.** En este caso se trata del aumento de las horas por cada año de aumento en la edad.

Recordemos que utilizamos la notación $\hat{y}(x)$ para indicar una ordenada obtenida a partir de la ecuación de la recta.

De la ecuación de la recta, estimamos una **media** poblacional **de 5,19 horas** dedicadas a actividades sedentarias entre todas las jóvenes de 15 años:

$$\begin{aligned}\hat{y}\ (\text{edad} = 15) &= 3{,}24 + 0{,}13 \text{ x } 15\\ &= 3{,}24 + 1{,}95\\ &= 5{,}19\end{aligned}$$

Para 16 años:

$$\begin{aligned}\hat{y}\ (\text{edad} = 16) &= 3{,}24 + 0{,}13 \text{ x } 16\\ &= 3{,}24 + 2{,}08\\ &= 5{,}32\end{aligned}$$

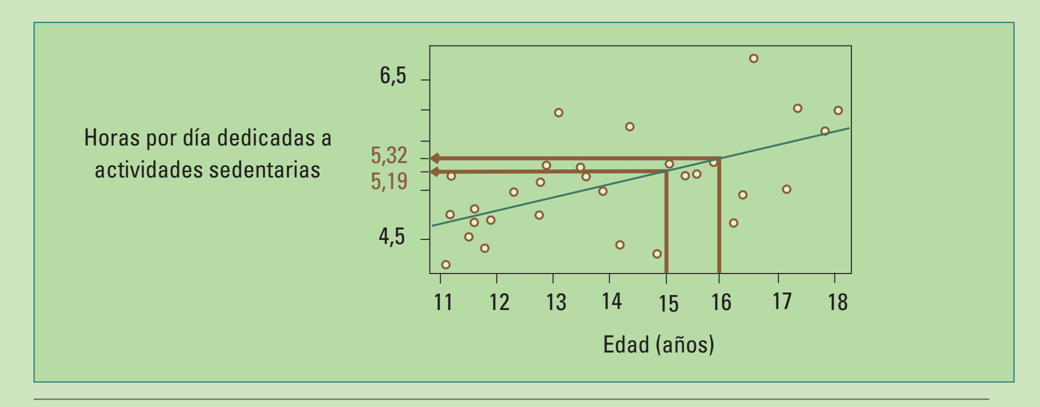


Figura 22.24. *Un aumento de un año corresponde a un aumento de 0,13 horas dedicadas a actividades sedentarias.*

Un aumento de un año en la variable explicativa corresponde a un aumento de 0,13 horas:

$$\hat{y}\,(\text{edad} = 16) = 5{,}32$$
$$\hat{y}\,(\text{edad} = 15) = 5{,}19$$
$$\hat{y}\,(\text{edad} = 16) - \hat{y}\,(\text{edad} = 15) = 5{,}32 - 5{,}19 = 0{,}13$$

Obtendríamos el mismo resultado al comparar 16 con 17 años, 12 con 13, etc. O sea, **por cada año de aumento en la edad** se espera **un aumento** de 7,8 minutos (= 0,13 x 60) en el tiempo dedicado a actividades sedentarias en las mujeres.

Este aumento (0,13) es la pendiente de la recta ajustada.

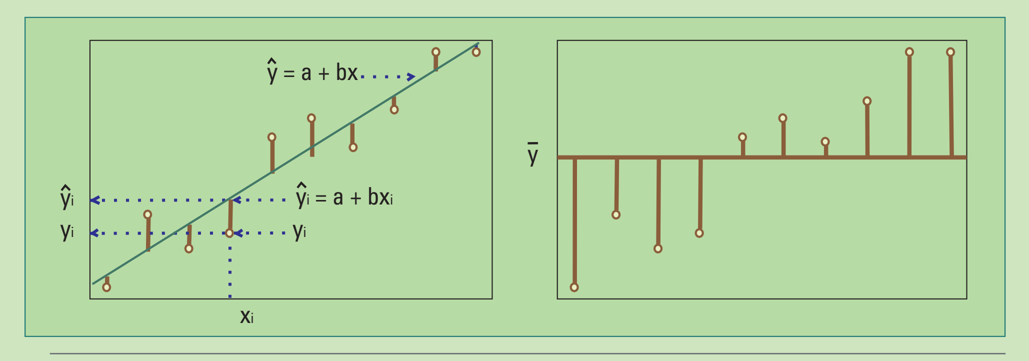
22.3.3.2. Interpretación de la recta ajustada. Coeficiente de determinación.

La recta de regresión ajustada en un diagrama de dispersión, y=a+bx , provee una estimación de la media poblacional de la variable respuesta en función de la variable explicativa.

> ! Recordemos que los coeficientes a y b de la recta estimada son números conocidos.
> En el ejemplo 6: y= 3,24 + 0,13 edad, por lo tanto a= 3,24 y b= 0,13

¿Qué significa $\hat{y}$ calculado como **a + bx** , para un valor determinado **x**? Es el **valor estimado** de la media poblacional de la variable respuesta para ese valor fijo de la variable explicativa.

¿Entonces $\hat{y}$ es la media muestral? No, pero juega un papel similar a la media muestral en el sentido de estimar una media poblacional, esta vez con un valor para cada valor de **x**.



22.25. *En el gráfico de la izquierda los residuos a la recta ajustada y=a+bx; a la derecha las distancias entre las respuestas observadas y su medida muestral ($\bar{y}$).*

> **!** ¿Por qué no tomamos directamente el promedio de los valores de la variable **Y** , y los promediamos? Porque la información de la variable explicativa contenida en los coeficientes **a** y **b** provee una estimación más precisa que la media muestral $\bar{y}$, en cuyo cálculo sólo se utilizan los valores de la variable **Y**.

Los valores observados ($\mathbf{y_i}$) están, en general más cerca de $\hat{\mathbf{y}}_i$ que de □**.** La figura 22.25 ilustra esta situación. Las líneas verticales rojas muestran cómo se desvían los valores observados de la variable respuesta ($\mathbf{y_i}$) respecto a la media muestral (□). Sus longitudes ($\mathbf{y_i}$ - □) al cuadrado son en promedio mayores que la de las líneas azules (□$_i$**-(a+bx$_i$)**)que miden las distancias de cada $\mathbf{y_i}$ a su correspondiente valor sobre la recta ajustada, $\hat{\mathbf{y}}_i$ Recordemos que cada cada $\mathbf{y_i}$ se obtiene un valor $\hat{\mathbf{y}}_i$ reemplazando $\mathbf{x_i}$ en la ecuación de la recta $\hat{\mathbf{y}}$ **= a+bx;** o sea $\hat{\mathbf{y}}_i$ **= a+bx$_i$.**

La figura 22.25 muestra como los valores observados ($\mathbf{y_i}$) tienen una distancia vertical a la recta ajustada ($\hat{\mathbf{y}}$ **= a+bx**) menor que la recta horizontal de ecuación **y =** □. El coeficiente de determinación, definido a continuación, cuantifica esa reducción.

El coeficiente de determinación ($\mathbf{R^2}$) mide la proporción en que se reduce la suma de las distancias verticales al cuadrado de los valores observados (**$\mathbf{x_i}$, $\mathbf{y_i}$**) alrededor de la recta de cuadrados mínimos en comparación con esas distancias alrededor de la recta de ecuación **y =** □**:**

$$R^2 = \frac{SCT - SCR}{SCT}$$

donde

- Suma de cuadrados total SCT. Mide las distancias verticales de los pares observados (x_i, y_i) a la recta de ecuación y = □

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- SCR es la suma de cuadrados de los residuos:

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

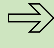
El coeficiente de correlación al cuadrado coincide con el coeficiente de determinación:

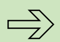
$$\mathbf{r^2 = R^2}$$

Ejemplo 8: Datos atletas. Cálculo del coeficiente de determinación.

DESARROLLO DEL CÁLCULO DEL COEFICIENTE DE DETERMINACIÓN. DATOS ATLETAS. TABLA 22.4

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	y_i - □	$(y_i - \hat{y}_i)^2$	(y_i - □$_i$)^2
1	47,17	47,83	46,42	1,41	4,13	1,99	17,06
2	47,41	46,03	46,31	-0,28	2,33	0,08	5,43
3	47,88	45,6	46,09	-0,49	1,9	0,24	3,61





i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$y_i - \bar{y}$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y}_i)^2$
4	51,05	48,55	44,61	3,94	4,85	15,52	23,52
5	51,32	42,37	44,48	-2,11	-1,33	4,45	1,77
6	52,18	43,93	44,08	-0,15	0,23	0,02	0,05
7	52,37	44,9	43,99	0,91	1,2	0,83	1,44
8	52,83	40,03	43,78	-3,75	-3,67	14,06	13,47
9	53,31	42,03	43,55	-1,52	-1,67	2,31	2,79
10	53,93	45,12	43,26	1,86	1,42	3,46	2,02
11	55,29	39,8	42,62	-2,82	-3,9	7,95	15,21
12	57,91	44,9	41,4	3,5	1,2	12,25	1,44
13	57,94	41,32	41,38	-0,06	-2,38	0,00	5,66
14	61,32	39,37	39,8	-0,43	-4,33	0,18	18,75
						63,36	**112,22**

$R^2 = 0{,}4354$

$\bar{y} = 43{,}699$

$$SCT = \sum_{i=1}^{14} (y_i - \bar{y})^2 = 112{,}22 \qquad \text{y} \qquad SCR = \sum_{i=1}^{14} (y_i - \hat{y}_i)^2 = 63{,}36$$

$$R^2 = \frac{112{,}22 - 63{,}36}{112{,}22} = 0{,}4354$$

La variabilidad de los datos de las atletas alrededor de la recta de regresión ajustada por cuadrados mínimos, y= 68,49 - 0,468 x, **se redujo en un 43,54%** en comparación con la variabilidad de la respuesta alrededor de su media muestral (recta y = 43,699).

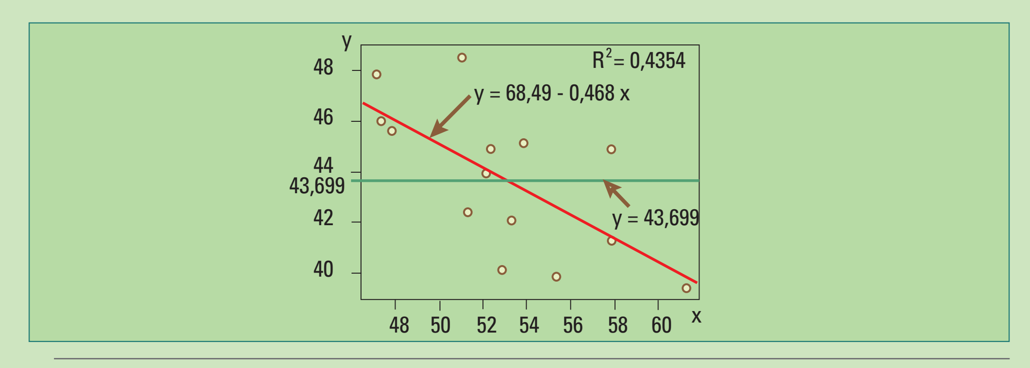


__Figura 22.26.__ Diagrama de dispersión datos atletas. En promedio las distancias a la recta de regresión son menores que a la recta y = 43,699.

Recordemos que r = -0,659, su cuadrado es 0,4342. La diferencia con R^2 se debe únicamente a errores de redondeo de los que no nos preocupamos porque es habitual que estos cálculos se realicen con muchos más decimales utilizando computadoras.

100 x R^2 es la **reducción de la variabilidad** de los valores (y_i) de la **variable respuesta** a una recta de regresión lineal y = a+bx en comparación con la variabilidad de los valores (y_i) respecto de la recta $y = \bar{y}$.

Decimos que **la recta de regresión explica** el 100 x R^2 de la variabilidad de la variable respuesta.

$R^2 = 0$ cuando la regresión no explica nada; en ese caso, la suma de cuadrados total es igual a la suma de cuadrados de los residuos.

$R^2 = 1$ cuando **todos los puntos están sobre la recta, la variabilidad** observada de la respuesta **es explicada totalmente por la regresión** y la suma de los cuadrados de los residuos es cero.

22.3.4 El método de cuadrados mínimos puede fallar

Alcanza con un solo dato "malo" para **arruinar completamente el resultado** de la **media muestral** y el **desvío estándar muestral.** Consideremos por ejemplo el siguiente conjunto de datos: 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16. Tiene media = 8 y desvío estándar = 5,05. Si el 16 es reemplazado por 1.600 (puede tratarse de un error o un valor de otra población) entonces el nuevo conjunto de datos tiene media = 101,18 y desvío estándar = 386,27.

En forma similar, **alcanza con un dato** "malo" para **arruinar completamente** a la recta de regresión al coeficiente de determinación y al coeficiente de correlación.

La figura 22.27 muestra dos diagramas de dispersión, los datos son los mismos en ambos diagramas salvo uno.

Para los datos del diagrama de la izquierda tenemos:

- Recta ajustada: **y = 0,99 + 1,99 x**
- Coeficiente de determinación: $\mathbf{R^2 = 0,9993}$
- Coeficiente de correlación: **r = 0,99965**

La recta ajustada tiene pendiente positiva y representa a la mayoría de los datos. El coeficiente de determinación es $R^2 = 0,9993$. Vimos su significado general en la sección 22.3.3.2., en este ejemplo dice que la recta explica el 99,93 % de la variabilidad de los datos. Un resultado similar se obtiene mediante el coeficiente de correlación r: su valor (0,99965) tan cercano a 1 refleja la casi linealidad de los puntos en el diagrama de dispersión.

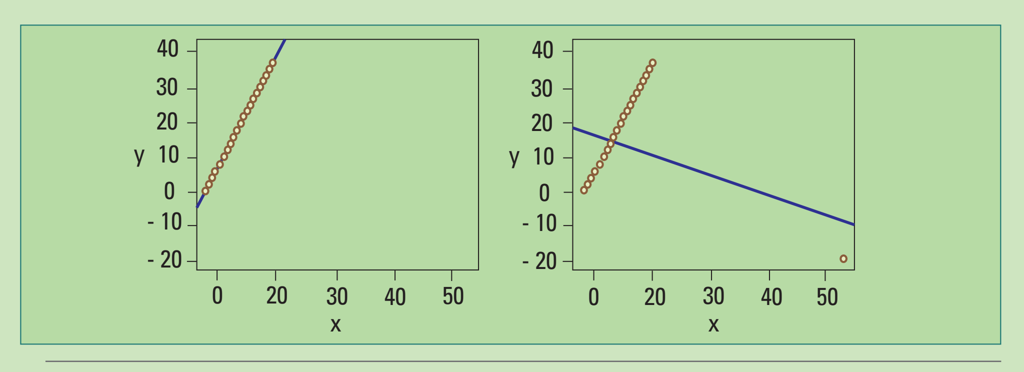
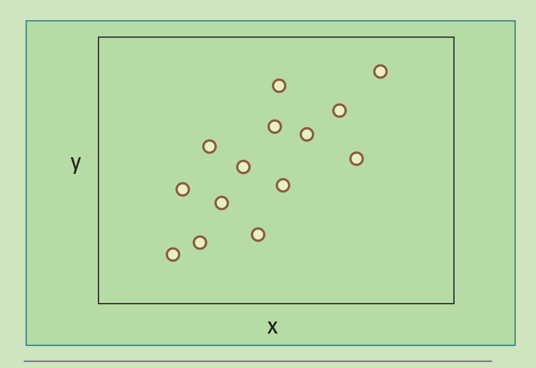


Figura 22.27. *Dos diagramas de dispersión y las rectas ajustadas. Un "punto palanca" distorsiona completamente el ajuste de la recta de cuadrados mínimos.*

Para los datos del diagrama de la derecha tenemos:

- Recta ajustada: **y = 17- 0,33 x**
- Coeficiente de determinación: $\mathbf{R^2}$ **= 0,09226**
- Coeficiente de correlación: **r = -0,30374**

Agregando un "punto palanca" el ajuste por cuadrados mínimos cambia completamente. Se trata de un punto que se encuentra alejado de la mayoría de los valores de la variable respuesta. La recta ajustada tiene pendiente negativa y su dirección es casi perpendicular a la de la mayoría de los datos. El coeficiente de correlación dice que la asociación entre las variables es negativa cuando en realidad, salvo por un dato la asociación es positiva.



22.28. *Diagrama de dispersión con forma de nube.*

El ajuste de una recta y el cálculo del coeficiente de correlación serán buenas medidas resumen de la relación entre dos variables continuas, siempre que el diagrama de dispersión tenga una forma de nube, como el de la figura 22.28, más o menos concentrada alrededor de una recta.

□ 22.4. Dos rectas

Muchas veces interesa comparar la relación de dos variables continuas en dos grupos distintos definidos por una variable categórica.

Ejemplo 9: Consideremos nuevamente los datos de la tabla 22.1. Esta vez, tendremos en cuenta la edad al comparar las horas por día dedicadas a las actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) entre varones y mujeres (tabla 22.1). Describiremos la relación entre la cantidad de horas y la edad en dos grupos; la variable categórica que los define es "género" y las categorías son: varón, mujer.

La figura 22.29 muestra el diagrama de dispersión de las horas por día dedicadas a actividades sedentarias para varones y mujeres y dos rectas de regresión. La recta azul fue estimada utilizando únicamente los datos de los varones y la roja los datos de las mujeres. Sus ecuaciones son:

- y = 2,84 + 0,197.edad, para varones
- y = 3,24 + 0,13.edad, para mujeres

Cada una de esas rectas estima la media en la población de la cantidad de horas que un varón (o una mujer) dedican a actividades sedentarias.

Sin tener en cuenta la edad, la diferencia de medias muestrales (5,73-5,06) de la cantidad de **horas** para varones y mujeres es **0,67 horas.**

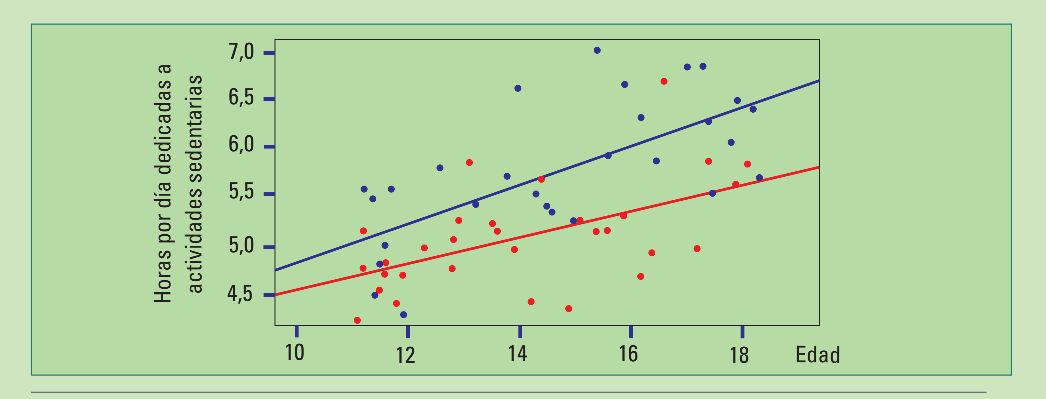


Figura 22.29. *Diagrama de dispersión de las horas por día dedicadas a actividades sedentarias en dos grupos: varones (en azul) y mujeres (en rojo) junto con sus respectivas rectas de regresión lineal simple.*

Teniendo en cuenta la edad. Las ecuaciones de las rectas de regresión simple, igual que la media muestral, proveen una estimación de la media poblacional, una para varones y otra para mujeres. Además esta vez se obtiene una estimación para cada valor de la variable explicativa (edad). Podemos estimar la misma diferencia que en el párrafo anterior para cada edad:

Media estimada de horas _ Varones = 2,84 + 0,197.edad
Media estimada de horas _ Mujeres = 3,24 + 0,13.edad

Diferencia de medias estimada = (2,84 + 0,197.edad) - (3,24 + 0.13 . edad)
= - 0,40 + 0,067.edad

Si edad = 12, resulta: Diferencia de medias estimada = -0,40 + 0,067 x 12
= 0,404

Si edad = 17, resulta: Diferencia de medias estimada = - 0,40 + 0,067 x 17
= 0,739

La figura 22.30 ilustra estas diferencias. Si edad = 12 obtenemos una estimación menor a 0,67 horas que se obtuvo simplemente restando las medias muestrales, si edad es 17 la diferencia es mayor.

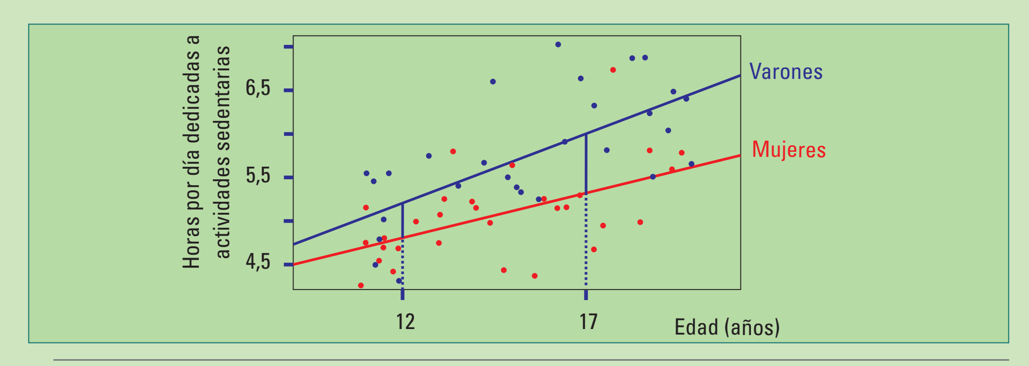


Figura 22.30. *A medida que aumenta la edad la diferencia de la cantidad de horas dedicadas a actividades sedentarias entre varones y mujeres también aumenta.*

☐ 22.5. Cuantificación de la relación entre dos variables categóricas

Muchas veces interesa estudiar la relación entre dos variables categóricas. En este caso no se puede usar la palabra "correlación" para describirla. La correlación es un caso especial de asociación, mide la fuerza de la relación lineal entre las variables numéricas. El término adecuado para variables categóricas es simplemente "asociación".

Consideremos, en general, dos variables categóricas con dos categorías cada una:

- "grupo de tratamiento" (sí, no)
- "resultado" (sí, no)"

Estarán asociadas si el porcentaje de pacientes con resultado (sí) en un grupo (sí) es muy diferente del porcentaje de pacientes con ese mismo resultado en el otro grupo (no).

Ejemplo 10: La tabla 22.5 muestra los resultados de una encuesta hipotética para determinar si existe una asociación entre comer rápido y el sobrepeso. Se obtuvieron los siguientes resultados:

Resultados de una encuesta hipotética para determinar si existe una asociación entre comer rápido y el sobrepeso. TABLA 22.5

		Come rápido	
		si	no
Sobrepeso	si	75	22
	no	175	198
		250	220
% con sobrepeso		30 =100x75/250	10 = 100x22/220

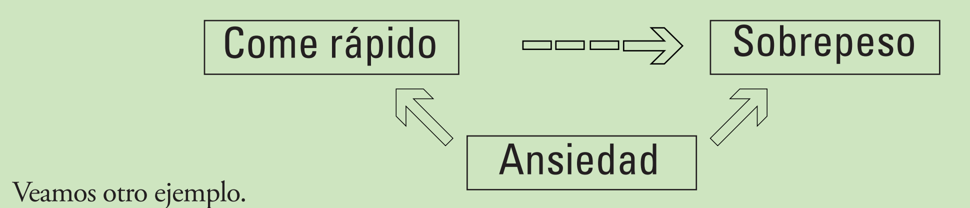
Las variables categóricas son:
Come rápido con categorías: si, no
Sobrepeso con categorías sí, no

En el grupo de individuos que come rápido el 30 % tiene sobrepeso; entre los que no comen rápido ese porcentaje se reduce al 10 %.

Los porcentajes de individuos con sobrepeso en las dos categorías de la variable come rápido son bastante diferentes, decimos que las dos variables están asociadas. En la sección 23.4.4 veremos si esa diferencia puede atribuirse a la variabilidad resultante del muestreo o es suficiente para establecer una asociación entre las variables.

22.6. Causalidad

Es muy tentador considerar una evidencia sobre asociación como una evidencia sobre causalidad. En el ejemplo 10 podríamos pensar que comer rápido es una causa del sobrepeso porque están asociados. Sin embargo, podría ocurrir que la ansiedad sea causante de comer rápido y también de comer demás y por lo tanto tener sobrepeso. La asociación entre comer rápido y sobrepeso es consecuencia de la ansiedad (una causa común a ambas).



Veamos otro ejemplo.

Ejemplo 11: Se realizó una encuesta para estudiar la relación entre la preferencia por el dulce de leche y el alfajor de dulce de leche. Se seleccionaron al azar 10 personas y se les solicitó que asignaran un número entre 0 y 100 a su preferencia. Donde 0 indica que a la persona no le gusta el dulce de leche o el alfajor y 100 que le apasiona.

La figura 22.31 muestra el diagrama de dispersión de los datos de la tabla 22.6. Vemos una clara tendencia lineal. Las personas que asignaron un puntaje alto al dulce de leche, también lo hicieron para el alfajor de dulce de leche. La gente no asignó los mismos puntajes en ambas escalas pero la preferencia por el dulce de leche es similar en rasgos generales a la del alfajor, aunque para este último un poco menor.

La figura 22.32 muestra los mismos datos junto con la recta de regresión lineal ajustada por cuadrados mínimos $y = -1,82 + 0,83\ x$. Cada vez que el puntaje del dulce de leche aumenta en 10 unidades, estimamos un aumento promedio de 8,3 unidades (10 x 0,83) para el puntaje del alfajor de dulce de leche.

PUNTAJE POR REFERENCIAS SOBRE EL DULCE DE LECHE Y EL ALFAJOR DE DULCE DE LECHE. TABLA 22.6

Persona	Dulce de leche	Alfajor de dulce de leche
1	100	85
2	90	80
3	60	44
4	60	50
5	20	25
6	95	70
7	90	80
8	70	50
9	50	30
10	10	5

Figura 22.31. *Diagrama de dispersión del puntaje asignado por preferencia al dulce de leche y al alfajor de dulce de leche, 0 indica que a la persona no le gusta y 100 que le apasiona.*

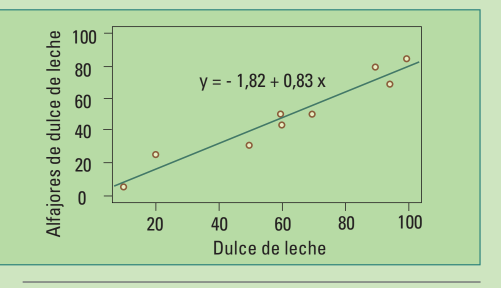


Figura 22.32. *Diagrama de dispersión del puntaje asignado por preferencia al dulce de leche y al alfajor de dulce de leche, junto con la recta de regresión lineal ajustada por cuadrados mínimos.*

El coeficiente de correlación **r** del puntaje (tabla 22.6) es 0,97. Este número es muy cercano a 1 (el valor máximo posible de **r**), mostrando que "la preferencia por el dulce de leche" y " la preferencia por el alfajor de dulce de leche" tienen un altísimo grado de asociación lineal. ¿Significa esto que la preferencia por el dulce de leche es la causa de la preferencia por el alfajor de dulce de leche? No necesariamente.

Las personas que le asignaron un puntaje alto al dulce de leche y al alfajor de dulce de leche pueden ser justamente las personas a las que les gusta la comida dulce y blanda. En este caso, la preferencia por ese tipo de comida sería la **causa común,** la que produce un alto grado de correlación entre las preferencias estudiadas.

En general, hallar una asociación entre variables es sólo un indicio. Si interesa establecer causalidad deberán realizarse estudios posteriores para confirmar o descartar las sospechas. Idealmente debería realizarse un experimento comparativo aleatorizado. La aleatorización produce grupos de sujetos, similares al comienzo de los tratamientos. Al comparar los grupos nos aseguramos que las diferencias observadas se deban a los efectos del tratamiento.

Pero muchas veces esto no es posible. Es difícil obligar a un sujeto a comer rápido o despacio. Sin embargo, sí es posible realizar un estudio observacional comparativo con grupos diferentes respecto del factor que se desea estimar y, lo más parecidos posibles en el resto. Si se sospecha que el nivel de ansiedad influye en el sobrepeso e interesa estudiar el efecto de comer rápido sobre el sobre peso, los grupos deberían ser similares en cuanto a nivel de ansiedad y también respecto de cualquier otra variable conocida que pueda afectar el resultado (edad, género, peso inicial, etc.).

□ 22.7. Más allá de un conjunto de datos

Cuando se ajusta una recta a pares de valores en un diagrama de dispersión el interés puede estar en, simplemente, obtener un resumen de la relación entre los puntos del diagrama, de la misma manera como puede interesar conocer un promedio o una proporción de un conjunto de datos. Este enfoque es llamado de **estadística descriptiva.** Se trata de obtener números que describen en forma resumida un conjunto de datos.

Pero si ese conjunto de datos proviene de un muestreo aleatorio simple de una población e **interesa** describir **el comportamiento de las variables en la población** (como ocurre la mayoría de las veces), la recta ajustada es una de las tantas rectas que se pueden obtener con diferentes muestras para el mismo problema. Se trata ahora de un problema de **inferencia estadística,** porque el interés ya no está centrado en el conjunto de datos que tenemos sino en toda la población de la cual provienen. En este caso **a, b** y **a+bx** son estimaciones de **α**, **ß**, y **α+ß** x respectivamente y **r** es una estimación de la correlación lineal de todos los valores las variables en la población. Como toda estimación tiene errores, por haber sido calculada a partir de una muestra en vez utilizar datos de toda la población. Se trata de **errores aleatorios debido al muestreo.**

Nos preguntamos qué pasaría si tomásemos muchas muestras de la misma población, ¿cómo cambiarían las rectas ajustadas?

Ya presentamos este enfoque en el capítulo 10 para proporciones cuando tratamos el "margen de error". En el capítulo 23, retomaremos el tema con más profundidad para medias muestrales y proporciones y daremos fórmulas de cálculo para los errores de estimación. Las fórmulas de cálculo de los errores de los coeficientes de la recta de regresión por cuadrados mínimos son matemáticamente más complejas, pero los conceptos estadísticos son similares a los que veremos en ese capítulo.

22.8. Actividades y ejercicios

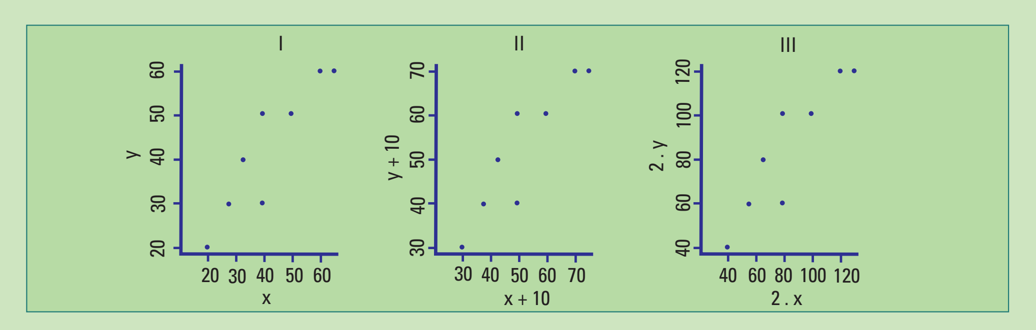
1.
a) ¿Cuáles son los todos los valores posibles del coeficiente de correlación?
b) ¿Cuáles son los todos los valores posibles del desvío estándar muestral **s**?

2. Muestre que

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

En los ejercicios 3 - 8 elija la respuesta correcta o la que completa la frase.

3. ¿Cuál de los siguientes tres diagramas de dispersión tiene mayor coeficiente de correlación?



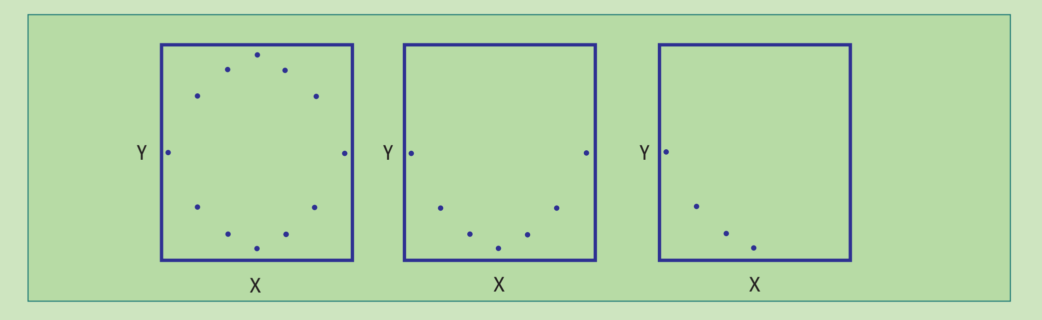
a) El I
b) El II
c) El III
d) Todos tienen el mismo coeficiente de correlación
e) No se puede responder a la pregunta porque falta información

4. Supongamos que el coeficiente de correlación es 0,7. Entonces, dados dos puntos del diagrama, ¿cuál de las siguientes situaciones es posible?

I) el punto que se encuentra a la izquierda (o sea con menor x) tiene un valor menor de y que el que se encuentra a la derecha.
II) el punto que se encuentra a la izquierda (o sea con menor x) tiene un valor mayor de y que el que se encuentra a la derecha.

a) Sólo I
b) Sólo II
c) I y II

5. ¿Cuál de las afirmaciones sobre los coeficientes de correlación de los puntos de los siguietes diagramas de dispersión son verdaderas?



a) Todos son cero.
b) Uno es cero, otro es positivo y otro es negativo.
c) Dos son cero y otro es cercano a -1.
d) Dos son cero y otro es -1.
d) Ninguno es cero.

6. Supongamos que la recta de regresión ajustada a un conjunto de datos **y** = 2+**bx** pasa-por el punto (3,11). Si $\bar{\mathbf{x}}$ e □ son las medias muestrales de los valores **x's** e **y's** respectivante, entonces □ =

a) $\bar{\mathbf{x}}$
b) $3\bar{\mathbf{x}}+2$
c) $\bar{\mathbf{x}}+2$
d) $2\bar{\mathbf{x}}-3$
e) $2\bar{\mathbf{x}}+3$

7. Un estudio determinó que el coeficiente de correlación entre el puntaje que los alumnos asignaron a sus profesores en una encuesta y el puntaje que la directora de la escuela les asignó a los mismos profesores es **r** = 1,25. Esto significa que

a) La directora y los alumnos coinciden en respecto a qué es un buen profesor.
b) La directora y los alumnos tienden a no estar de acuerdo respecto a qué es un buen profesor.
c) Hay poca relación entre los puntajes.
d) La asociación entre ambos puntajes es fuerte.
e) Hay un error de cálculo

8. El coeficiente de correlación satisface:

I. No es afectado por cambios en las unidades en que se miden las variables.

II. No es afectado por intercambiar las variables que se ponen en x e y.
III. No es afectado por la presencia de valores atípicos.

a) I y II
b) I y III
c) II y III
d) I, II y III
e) Ninguna de las afirmaciones es correcta.

9. ¿Le dedican los varones de su escuela más horas a realizar actividades sedentarias que las mujeres? Realice una encuesta para responder esta pregunta.

23. Teorema central del límite (TCL)

Uno de los resultados más importantes de la teoría estadística.

En algunas ocasiones nos interesa saber, por ejemplo, cuál es el peso medio de los recién nacidos, o cuál es la proporción de alumnos que no entienden estadística, sin embargo, en realidad no lo podemos saber exactamente. Aún cuando se recolecte con cuidado una muestra tan grande y representativa como sea posible, el error debido al muestreo es inevitable.

Nos interesa conocer los **parámetros poblacionales,** la media μ_X y el desvío σ_X estándar de una variable X en la población (es decir de todos los valores que toma la variable en la población); pero sólo podemos hallar la **media muestral** ($\bar{x}$) y el desvío estándar muestral (s). Se trata de **estadísticos** cuyos valores cambian entre una muestra y otra.

El resultado descripto en este capítulo permite cuantificar el error debido al muestreo. Comenzamos con preguntar: ¿qué pasaría si tomásemos muchas muestras de la misma población?

Ya planteamos esa pregunta en el capítulo 10 para proporciones, ahora consideraremos primero la media muestral y luego volveremos con las proporciones.

□ 23.1. Distribución de muestreo de la media muestral

Describiremos la **distribución de muestreo** de la **media muestral.** ¿Qué significa esto? ¿Cuál es la población? La **población** de muestreo de la media muestral está compuesta por **todas las medias muestrales** que se podrían obtener de una población utilizando **muestreos aleatorios simples,** observando los valores de alguna variable (**X**) y calculando la media muestral para cada una de los muestreos realizados. La distribución de muestreo de la media muestral tiene, como cualquier otra distribución, forma, centro y una medida de variabilidad. Conocer la distribución de muestreo de la media muestral permite estimar su error.

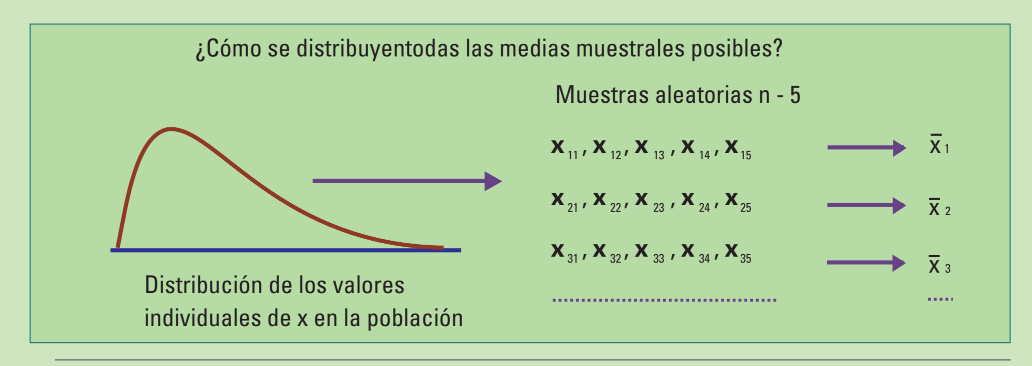


Figura 23.1. *Repetición del muestreo aleatorio simple de tamaño n=5.*

Muchas veces diremos: **"distribución de $\bar{x}$"**, sobreentendiendo que se trata de la **distribución de las medias muestrales en un muestreo aleatorio simple** de muestras de tamaño **n.**

Para tener una idea más concreta acerca de cómo se compara la distribución de una variable (**X**) en la población y la distribución de muestreo de $\bar{\mathbf{x}}$, veamos un caso particular de cómo se distribuyen 20 medias muestrales, calculadas a partir de muestras de tamaño n=4 y todos los valores individuales.

Ejemplo 1: volvamos a la fábrica que produce **garrafas de gas comprimido** de uso doméstico con 20 kg de capacidad nominal (sección 21.1.1.1), pero esta vez, para estudiar como se comporta la **media muestral.** Seleccionamos una muestra de 4 garrafas de un lote grande y calculamos la media de los volúmenes obtenidos. Repetimos este procedimiento 20 veces:

¿Cómo se distribuyen los volúmenes de las garrafas seleccionadas? ¿Cómo se distribuyen las medias muestrales?

Los valores individuales se encuentran entre 45,37 y 48,48 (dentro de un intervalo de longitud 3,19), las medias están entre 46,59 y 47,42 (ahora, el intervalo es de longitud menor a 1).

VOLÚMENES (dm^3) DE 80 GARRAFAS DE 20 kg.

TABLA 23.1

					Media muestral $\bar{x}$
Muestra 1	46.40	47.62	48.39	47.75	47.54
Muestra 2	47.08	46.39	45.37	48.08	46.73
Muestra 3	47.05	47.54	47.34	47.79	47.43
Muestra 4	46.68	47.51	46.89	47.39	47.12
Muestra 5	46.91	48.00	47.97	46.82	47.42
Muestra 6	47.29	46.73	46.47	47.03	46.88
Muestra 7	46.92	47.07	47.23	47.36	47.14
Muestra 8	47.76	46.88	48.16	47.26	47.51
Muestra 9	46.30	47.76	47.18	46.65	46.97
Muestra 10	46.79	45.83	45.70	46.75	46.27
Muestra 11	47.08	45.57	46.45	48.48	46.90
Muestra 12	46.57	46.54	46.98	46.35	46.61
Muestra 13	48.37	48.04	46.65	46.01	47.27
Muestra 14	47.61	47.02	47.44	46.87	47.24
Muestra 15	46.92	47.52	46.80	46.50	46.94
Muestra 16	47.07	47.90	46.33	46.46	46.94
Muestra 17	46.76	47.17	47.33	46.34	46.90
Muestra 18	47.75	47.39	47.14	46.18	47.12
Muestra 19	46.67	47.04	48.15	47.50	47.34
Muestra 20	46.28	47.12	46.23	47.52	46.79

Para tener una imagen visual construimos dos histogramas en escala densidad (Recordemos que en escala densidad el área de cada rectángulo coincide con la proporción de datos que pertenecen al correspondiente intervalo de clase, sección 17.2). El primero con todos los datos individuales (columnas 2, 3, 4 y 5 de la tabla 23.1) y el segundo con las medias muestrales (columna 6). Utilizamos las mismas escalas en los ejes de ambos histogramas para poder compararlos.

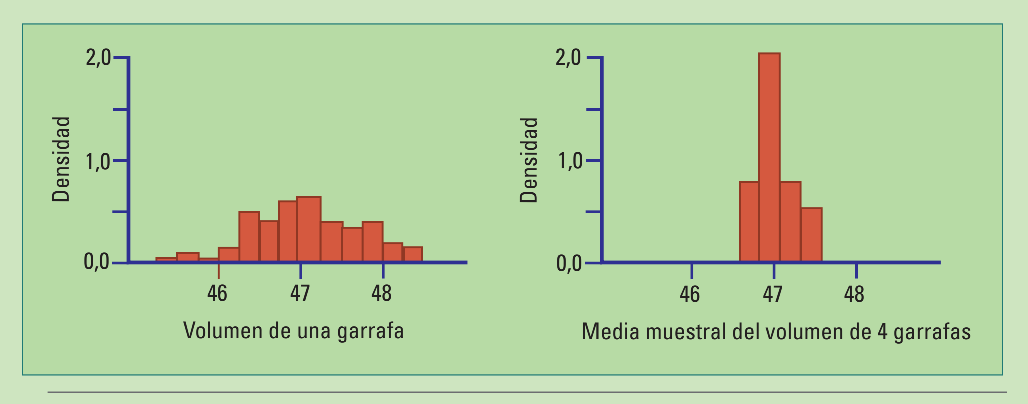
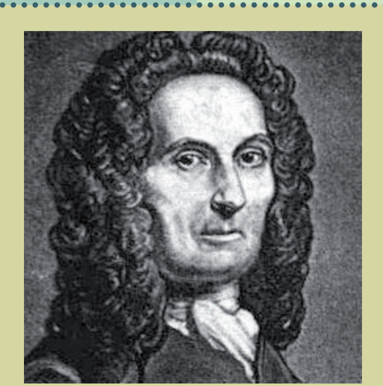


Figura 23.2. *Histogramas de datos de volumen de garrafas de 20 kg. Las medias muestrales (histograma lado derecho) están más concentradas alrededor del valor central 47 que los valores individuales (histograma lado izquierdo).*



Abraham de Moivre 1667 - 1754. Matemático francés.

En 1733, postuló la primera versión del Teorema Central del Límite en el marco de los juegos de azar. Específicamente, lo hizo como descripción de la distribución de "la cantidad de caras" que resultan al arrojar "una moneda equilibrada" muchas veces.

Muchos años después (1812), el tema fue retomado, generalizado a la "cantidad de caras de cualquier moneda" (aunque sea una moneda cargada) y demostrado por otro matemático francés: Pierre-Simon Laplace.

Las medias muestrales tienen menor variabilidad que los valores individuales.

¡Por eso promediamos!

Le creemos más al resultado de un promedio que a un resultado individual.

Además de tener menor variabilidad, la distribución de las medias muestrales tiene otra propiedad muy importante. Esta propiedad la establece el Teorema Central del Límite.

23.2. Enunciado del TCL

Supongamos que una variable se distribuye en la población con media μ y desvío estándar σ. Si se realizan muestreos aleatorios simples y se registran los valores $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ de dicha variable entonces:

- La distribución de muestreo de $\bar{x}$, es **aproximadamente Normal** con tal de tomar tamaños de muestra suficientemente grandes.
- Cuanto mayor sea el tamaño de la muestra (**n**) tanto mejor será la aproximación. Si **n** es por lo menos 30, la aproximación será buena en la mayoría de los casos.
- La media de la distribución de $\bar{x}$ también es **μ.**
- El desvío estándar de la distribución de $\bar{x}$, es $\frac{\sigma}{\sqrt{n}}$. Decrece cuando aumenta el tamaño de la muestra (**n**).
- La distribución de muestreo de $\sqrt{n}\,\frac{\bar{x}-\mu}{\sigma}$, es **aproximadamente Normal Estándar** siempre que se utilicen tamaños de muestra suficientemente grandes

El TCL puede enunciarse, bajo los mismos supuestos, para la suma:

- La distribución de muestreo de $\sum_{i=1}^{n} x_i$, es **aproximadamente Normal** siempre que se utilicen tamaños de muestra suficientemente grandes.
- Cuanto mayor sea el tamaño de la muestra (**n**) tanto mejor será la aproximación. Si **n** es por lo menos 30, la aproximación será buena en la mayoría de los casos.
- La media de la distribución de $\sum_{i=1}^{n} x_i$ también es n **μ.**
- El desvío estándar de la distribución de $\sum_{i=1}^{n} x_i$, es $\sigma\sqrt{n}$.

Observación. Como $\sum_{i=1}^{n} x_i = n\bar{x}$, los valores de muestreo de $\sum_{i=1}^{n} x_i$ y $\bar{\mathbf{x}}$ difieren únicamente en un factor multiplicativo (**n**). Por lo tanto, la forma de sus distribuciones es la misma; aproximadamente Normal aunque con diferente media y diferente desvío estándar.

Frecuentemente no se conoce σ, entonces se lo estima por **s** el desvío estándar de la muestra (sección 18.2.3):

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

El desvío estándar estimado de la distribución de $\bar{\mathbf{x}}$ utilizando **s** se denomina error estándar.

- Se denomina **error estándar** a $\frac{s}{\sqrt{n}}$ y se lo utiliza para estimar $\frac{\sigma}{\sqrt{n}}$.

El TCL, además de mostrar cómo la media muestral da una estimación más precisa cuanto más grande sea la muestra de la cual proviene, con un desvío estándar de $\frac{\sigma}{\sqrt{n}}$ describe que esas medias muestrales se distribuyen en forma **aproximadamente Normal,** cualquiera sea la forma de la distribución de los datos individuales.

Generalmente, se considera que **n=30** es un tamaño de muestra adecuado para que la distribución de $\bar{x}$ pueda aproximarse por la Normal. En realidad, ese tamaño puede ser bastante menor si la distribución de los datos individuales es simétrica, y debe ser mayor si esa la distribución es muy diferente de la Normal.

Si los datos provienen de una variable que tiene distribución Normal en la población, $\bar{x}$ tendrá una **distribución Normal, cualquiera sea n.**

Si los datos provienen de una variable que **no tiene** distribución Normal en la población, $\bar{x}$ tendrá distribución a**proximadamente Normal** siempre que **n sea grande.**

Siguiendo con el ejemplo de las garrafas, supongamos que los volúmenes, de **todas las posibles garrafas producidas de la misma forma,** se distribuyen de forma Normal con media **μ=47** dm^3 y desvío **σ=0,75** dm^3. La figura 23.3 muestra la distribución de $\bar{x}$ para una muestra de tamaño 4 junto con la distribución de las observaciones individuales. El desvío estándar de la distribución de $\bar{x}$ con **n=4** es la mitad que el desvío estándar de distribución de las observaciones individuales.

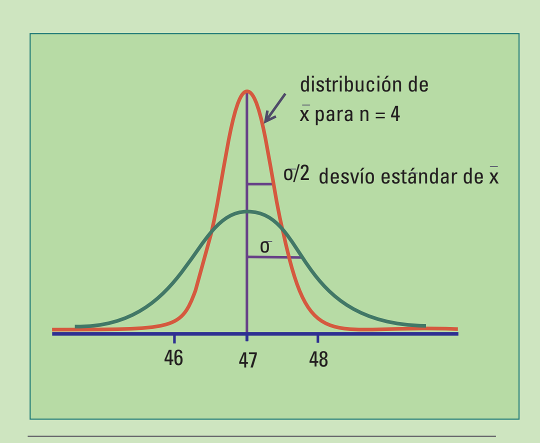


Figura 23.3. *Distribución de $\bar{x}$ para un muestreo aleatorio simple con n=4, junto con la distribución de las observaciones individuales.*

Ejemplo 2: La distribución de la longitud de las palabras puede distinguir idiomas, estilos de escritura e incluso hasta autores. La figura 23.4 muestra la distribución presentada en la resolución del ejercicio 4 de la sección 17.3. Supondremos que esta es efectivamente la distribución de la longitud de las palabras de todo un libro. Esas longitudes tienen media **μ=5,51** letras por palabra y desvío estándar **σ=3,56**, valores que en lo que sigue, se consideran parámetros poblacionales. A partir de ahora, pensamos en una "población de palabras" con media **μ=5,51** letras por palabra y desvío estándar σ=**3,56.**

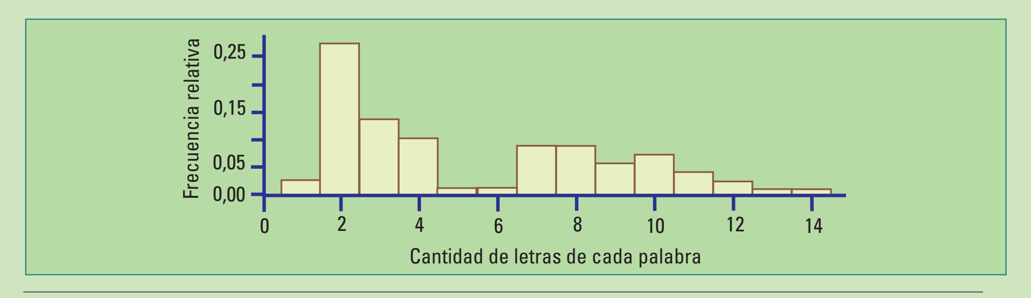


Figura 23.4. *Distribución de la cantidad de letras por palabra de un libro en castellano.*

Supongamos que elegimos al azar 1.000 muestras de 3 palabras de ese libro, registramos la cantidad de letras que tiene cada una de las palabras elegidas y calculamos la media de cada una de las 1.000 muestras. Tendremos así 1.000 números cuyo histograma mostrará su distribución. Repetimos el mismo procedimiento anterior pero con muestras de tamaño 30. Las figuras siguientes muestran los resultados:

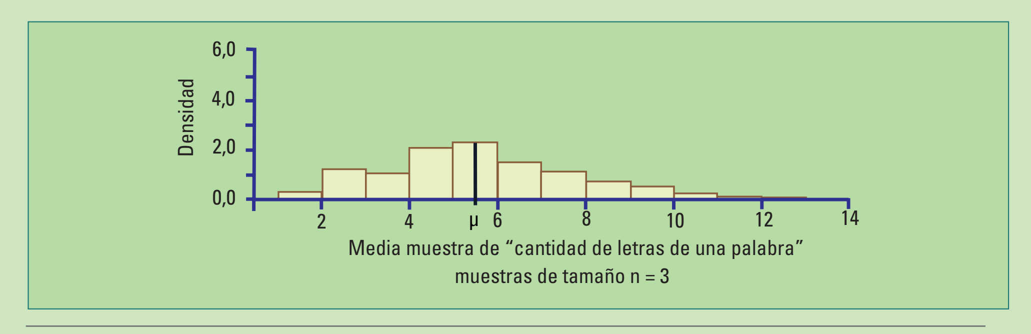


Figura 23.5. *Distribución de la media muestral para tamaño de muestra 3.*

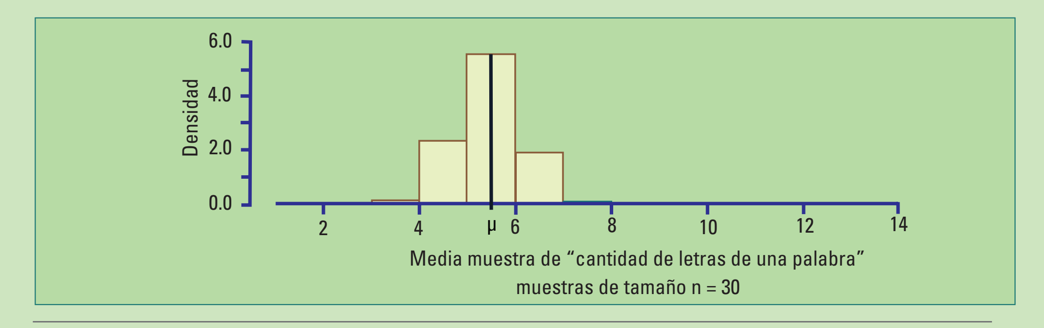


Figura 23.6. *Distribución de la media muestral para tamaño de muestra 30.*

Los valores de las medias muestrales son más simétricos y están más concentrados alrededor de la media poblacional (**μ=5,51**) cuando se toman muestras de tamaño 30, que cuando solamente las muestras tienen 3 palabras cada una (figuras 23.5 y 23.6).





En los ejemplos 1 y 2 se generaron muchas medias muestrales para conocer cómo se distribuyen.

En la práctica, la media muestral se calcula una única vez y no conocemos la distribución poblacional.

Entonces, ¿para qué sirve lo anterior?

Sirve para:

- Ver que las medias muestrales obtenidas con más datos, en general estarán más cerca de la media poblacional aunque no la conozcamos.
- Conocer la distribución y determinar un rango de valores donde se pueden encontrar la mayoría de las medias muestrales, **aunque no conozcamos la distribución poblacional** de la que provienen los datos.

◻ 23.3. Distribución de muestreo de la proporción muestral

El TCL no se aplica sólo a la media muestral de una variable numérica (ejemplos 1 y 2), también se utiliza para la proporción muestral $\hat{\mathbf{p}}$ (se lee p-sombrero). Se trata, en este caso, de un análisis cualitativo observando la presencia o ausencia de una característica para estimar su proporción p en la población. La población se divide en dos partes, los que tienen y los que no tienen la característica en cuestión.

Vimos esta situación con el ejemplo del "Club Grande de Fútbol":

- En el capítulo 9 la investigadora tomó una **única muestra** de 538 socios, para cada uno se observó si estaba a favor (presencia) o no estaba a favor (ausencia) del candidato 1; obteniendo $\hat{\mathbf{p}}$ **=0,51.**
- En el capítulo 10 se presentaron los resultados de seleccionar **1.000 muestras** aleatorias simples, cada una de **tamaño 538,** de una población para la cual la verdadera proporción es **p=0,5.**

La **distribución de muestreo** de $\hat{\mathbf{p}}$ es la distribución de **todas** las **proporciones muestrales** posibles, calculadas a partir de muestras aleatorias simples del mismo tamaño **n.**

Diremos: **"distribución de $\hat{p}$"**, sobreentendiendo que se trata de la **distribución de la proporción muestral en un muestreo aleatorio simple.**

Para cualquier población que tiene una proporción **p** de individuos con una característica (éxito) y

$$\hat{p} = \frac{\text{cantidad de éxitos en la muestra}}{\text{tamaño de la muestra}}$$

es la proporción de éxitos en una muestra aleatoria simple de **n** individuos, entonces el TCL asegura que:

- La **media** de la distribución de $\hat{p}$ es **p**, ($\mu_{\hat{p}} = p$).
- El **desvío estándar** de la distribución de $\hat{p}$ es $\sqrt{\frac{p(1-p)}{n}}$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- La distribución de $\hat{p}$ es **aproximadamente Normal,** siempre que tanto n p y n (1 - p) sean grandes (mayores a 10) y la muestra sea una muestra aleatoria simple.
- La distribución de muestreo de $\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}}$, es **aproximadamente Normal Estándar** siempre
- que, **n p** y **n (1-p)** sean grandes (mayores a 10) y la muestra sea una muestra aleatoria simple.

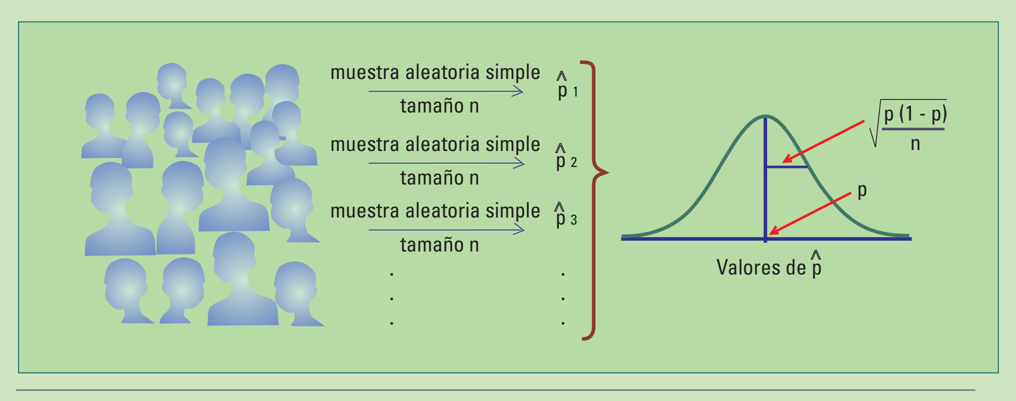


Figura 23.7. *Repetición del muestreo aleatorio simple de tamaño n. Los valores de $\hat{p}$ tienen distribución Normal con media p y desvío estándar* $= \sqrt{\frac{p(1-p)}{n}}$

Como **p** es desconocida, una estimación razonable es utilizar $\hat{\mathbf{p}}$ para estimar el desvío estándar, esta estimación se denomina error estándar:

$$\textbf{Error estándar} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

□ 23.4. Corrección por tamaño de población

Hemos realizado muestreos aleatorios simples sin tener en cuenta el tamaño de la población (**N**), esto es válido siempre que el **tamaño de la muestra** (**n**) sea sólo **una pequeña parte de la población.** Ya utilizamos esta propiedad al introducir el concepto de margen de error en la sección 10.2.

Cuando **la muestra no es una pequeña parte** de la **población,**

- la media de la distribución de muestreo de $\hat{p}$ ($\mu_{\hat{p}}$) sigue siendo p, pero
-
- el desvío estándar de la distribución de muestreo ($\sigma_{\hat{p}}$) es $\sqrt{\frac{N-n}{N-1}}\ \sqrt{\frac{p\,(1-p)}{n}}$

El **factor de corrección** $\sqrt{\frac{N-n}{N-1}}$ que multiplica a $\sqrt{\frac{p\,(1-p)}{n}}$ (el desvío estándar de $\hat{p}$ dado en la sección 23.3) para corregir el desvío estándar **es el mismo** que debe utilizarse para el **error estándar** y el **margen de error.**

En el ejemplo del Club Grande de Fútbol, **N=58.210** y **n=538**, **no es necesario utilizar el factor de corrección** ya que es muy cercano a 1:

$$\sqrt{\frac{N-n}{N-1}}\ \sqrt{\frac{58.210 - 538}{58.210 - 2}} = 0{,}996$$

DISTRIBUCIÓN DE MUESTREO DE $\hat{P}$ PARA n=3 Y N=5. TABLA 23.2

		Proporción de alumnos que no estudió
Muestra 1	ABC	$\hat{p}_1 = 0$
Muestra 2	ABD	$\hat{p}_2 = 1/3$
Muestra 3	ABE	$\hat{p}_3 = 1/3$
Muestra 4	ACD	$\hat{p}_4 = 1/3$
Muestra 5	ACE	$\hat{p}_5 = 1/3$
Muestra 6	ADE	$\hat{p}_6 = 2/3$
Muestra 7	BCD	$\hat{p}_7 = 1/3$
Muestra 8	BCE	$\hat{p}_8 = 1/3$
Muestra 9	BDE	$\hat{p}_9 = 2/3$
Muestra 10	CDE	$\hat{p}_{10} = 2/3$

El siguiente es un ejemplo hipotético. El tamaño de la muestra es más de la mitad de la población. La población es conocida y pequeña para que los cálculos sean razonablemente cortos. **No vale la aproximación de la distribución de muestreo** de $\hat{p}$ por la Normal, pero sí las conclusiones respecto de su media y su desvío estándar.

Ejemplo 3: Una población está formada por cinco alumnos (**N=5**) A, B, C, D, E, que van a rendir examen de estadística, en la que 2 de ellos (**D** y **E**) no estudiaron. La proporción de alumnos que no estudió es **p=2/5=0,4.** Se eligen 3 alumnos al azar para saber qué proporción de alumnos estudió.

La tabla 23.2 muestra la lista de las **10 muestras posibles** de tamaño **n=3** de la población y la proporción muestral de alumnos que no estudió.

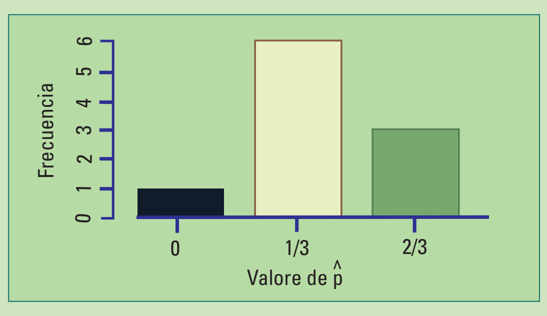


Figura 23.8. *Distribución de muestreo de $\hat{p}$, datos de la tabla 23.2.*

Utilizamos los valores de la tabla 23.2 para calcular la **media** de la distribución de muestreo de $\hat{\mathbf{p}}$ (promedio de **todos** los **valores posibles** de $\hat{\mathbf{p}}$):

$$\mu_{\hat{p}} = \frac{\sum_{i=1}^{10} \hat{p}_i}{10}$$

$$\mu_{\hat{p}} = \frac{0+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{2}{3}+\frac{1}{3}+\frac{1}{3}+\frac{2}{3}+\frac{2}{3}}{10}$$

$$\mu_{\hat{p}} = \frac{6\times\frac{1}{3}+3\times\frac{2}{3}}{10} = \frac{4}{10} = 0{,}4$$

$\mu_{\hat{p}}$ es un parámetro poblacional (de la distribución de muestreo de $\hat{\mathbf{p}}$) y coincide con la proporción poblacional p=0,4.

Ahora, calculamos el desvío estándar de los valores de $\hat{\mathbf{p}}$, $\sigma_{\hat{p}}$.

Calculemos el desvío estándar poblacional (es un parámetro) para los valores de $\hat{\mathbf{p}}$:

$$\sigma_{\hat{p}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i-\mu_{\hat{p}})^2} = \sqrt{\frac{1}{10} \sum_{i=1}^{10} (\hat{p}_i-0{,}4)^2} = 0{,}2$$

No debe preocupar el cálculo anterior, las calculadoras lo hacen bien.

Llama la atención que $\sigma_{\hat{p}}$ **no es** $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}4\ (1-0{,}4)}{3}} = 0{,}2828$ (el desvío es tándar de $\hat{\mathbf{p}}$ dado en la sección 23.3), sino un número más pequeño (0,2) que puede obtenerse mediante el siguiente cálculo:

$$\sigma_{\hat{p}} = \sqrt{\frac{N-n}{N-1}}\ \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{5-3}{5-1}}\ \sqrt{\frac{0{,}4(1-0{,}4)}{3}} = 0{,}2$$

$\sqrt{\frac{5-3}{5-1}} = 0{,}7071$ es el factor de corrección en este ejemplo.

En general:

El factor de corrección es $\sqrt{\frac{N-n}{N-1}}$

Lo anterior es para proporciones, pero ¿qué ocurre con la media muestral cuando la muestra no es una pequeña parte de la población?

- La media de la distribución de $\bar{\mathbf{x}}$ ($\mu_{\bar{X}}$) sigue siendo μ, pero
-
- El desvío estándar de la distribución de $\bar{\mathbf{x}}$ ($\sigma_{\bar{X}}$) es $\sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$

El mismo factor de corrección se utiliza corregir el desvío estándar de la distribución de $\hat{\mathbf{p}}$ y de $\bar{\mathbf{x}}$.

El factor de corrección es:

- Menor o igual a 1, siempre.
- Cercano a 1, si el tamaño de la población (**N**) es muy grande comparado con el tamaño de la muestra (**n**) y no es necesario utilizarlo.
- Cercano a 0, si el tamaño de la muestra se acerca al tamaño de la población, (**n≈ N**).
- Cero, si **n = N.**

Por lo tanto, como era de esperar:

Si la muestra es toda la población, el **desvío estándar** de la distribución de $\bar{\mathbf{p}}$ y de $\bar{\mathbf{x}}$**,** y **el margen de error valen** 0.

□ 23.5. El TCL y el mundo real

El Teorema Central del Límite sirve para interpretar los resultados de una encuesta o un estudio.

Permite aplicar, tanto para la **media muestral** como para la **proporción muestral,** las propiedades que vimos (sección 20.2.3) sobre datos que se distribuyen de acuerdo con la curva Normal; siempre que el tamaño de la muestra sea suficientemente grande. Por supuesto, las observaciones deben ser mediciones válidas y no tener sesgo.

Por ejemplo, cuando se trata de medias muestrales cuya distribución tienen media μ y desvío estándar $\frac{\sigma}{\sqrt{n}}$, permite afirmar que:

- Cerca del 95% de las medias muestrales están entre μ menos 2 veces su desvío estándar, y µ más 2 veces su desvío estándar, o sea dentro de $\mu \pm 2\frac{\sigma}{\sqrt{n}}$.
- El 99,7% (casi todas) de las medias muestrales se encuentran en el intervalo ($\mu - 3\frac{\sigma}{\sqrt{n}}$, $\mu + 3\frac{\sigma}{\sqrt{n}}$), o sea dentro de $\mu \pm 3\frac{\sigma}{\sqrt{n}}$.

Ya hemos utilizado este último resultado al establecer los límites de control en un gráfico equis barra (sección 21.2.2):

- Límite inferior de control (LIC) = $\mu - 3\frac{\sigma}{\sqrt{n}}$
- Límite superior de control (LSC) = $\mu + 3\frac{\sigma}{\sqrt{n}}$

De esta manera si el proceso está en control, solamente 0,3 % de las veces una media muestral excederá los límites dando una alarma falsa.

El Teorema Central del Límite se utiliza también como argumento para tratar de explicar porqué frecuentemente los errores de medición, y muchas variables relacionadas con fenómenos naturales, tienen distribuciones que pueden aproximarse por curvas de Gauss.

Por ejemplo, podemos pensar el error de medición como una suma de pequeños errores, luego por el TCL, es razonable que estos errores tengan distribución Normal. La idea puede aplicarse también a cualquier variable que se pueda considerar como una suma de pequeñas contribuciones independientes (no sirve sumar siempre el mismo número). Generalizaciones del teorema central del límite a promedios pesados y a sumandos parcialmente independientes muestran cuándo se obtiene una buena aproximación a la Normal y cuando no.

Desde que Abraham de Moivre postuló la primera versión del TCL en 1733, muchos matemáticos célebres -Pierre Simon, marqués de Laplace (1812), Siméon Denis Poisson (1824), Pafnuty Tchebyshev (1887), Aleksandr Lyapunov (1901), Jarl Waldemar Lindeberg (1922), para mencionar algunos de los más famosos- obtuvieron resultados respecto a la aproximación por la Normal a la distribución de medias o sumas. Más que un único teorema, se trata de muchos dependiendo de las condiciones que se requieren para su validez. Estas generalizaciones siguen siendo aún temas de investigación en estadística. Las demostraciones utilizan herramientas matemáticas avanzadas y resultados de la teoría de probabilidad.

Utilizaremos los resultados del TCL en los capítulos 24 y 25 en la construcción de dos nuevas herramientas estadísticas.

23.6. Actividades y ejercicios

1. Indique cuáles de las siguientes afirmaciones son verdaderas y justifique brevemente.

   a) Cuanto mayor sea el tamaño de la muestra mayor será el desvío estándar de la distribución de muestreo de $\bar{\mathbf{x}}$.
   b) El desvío estándar de la distribución de muestreo de $\bar{\mathbf{x}}$ sólo depende del tamaño de la muestra.
   c) La distribución de muestreo de $\bar{\mathbf{x}}$ es Normal si la población tiene una distribución Normal.
   d) Cuando n es grande la distribución de muestreo de $\bar{\mathbf{x}}$ es aproximadamente Normal, aún cuando la población no tenga una distribución Normal.

2. Indique cuales de las siguientes afirmaciones son verdaderas y justifique brevemente.

   a) La distribución de muestreo de $\hat{\mathbf{p}}$ tiene media igual a la proporción poblacional **p**.
   b) La distribución de muestreo de $\hat{\mathbf{p}}$ tiene un desvío estándar igual a $\sqrt{np(1-p)}$.
   c) Se considera que la distribución de muestreo de $\hat{\mathbf{p}}$ es aproximadamente Normal cuando $\mathbf{n \geq 30}$.

3. Realice los siguientes experimentos, puede pedir ayuda a sus compañeros y realizar una repetición cada uno:

   a) Arroje un dado 3 veces, si sale 1 ó 2 registre éxito, en caso contrario registre fracaso. Cuente la cantidad de éxitos. Repita 60 veces.
   b) Arroje un dado 30 veces, si sale 1 ó 2 registre éxito, en caso contrario registre fracaso.

   Para cada uno de los experimentos anteriores construya la tabla siguiente:

Número de repetición	Cantidad de éxitos	Proporción de éxitos
1		
2		
3		
.		
.		
60		

Halle el histograma de las proporciones de éxitos obtenidas. Compare.

En la práctica, la distribución de los valores de una variable en la poblción no cambia si se quita de ella una proporción muy pequeña. Los experimentos anteriores representan los resultados de muestreos aleatorios simples con reposición (o sea, la población no cambia con cada extracción) para estimar la proporción de éxitos (**p**).

En el ejemplo del dado, se registra éxito cuando al arrojarlo sale 1 ó 2. Si el dado está equilibrado, 2 resultados entre 6 posibles dan éxito en una proporción **p=1/3**. Los resultados de arrojar el dado representan muestreos aleatorios simples con reposición de una población cuya proporción de "éxitos" es 1/3. En general, no conocemos la verdadera proporción de "éxitos" en una población real.

4. Interesa saber cómo se comporta la media y el desvío de la distribución de muestreo de $\bar{\mathbf{x}}$ cuando el tamaño de la muestra es una parte importante del tamaño de la población.

   Considere una población de tamaño **N=6** y una variable en la población toma los valores 2 4 6 9 12 18.

   a) Calcule la media y el desvío poblacional de la variable ($\boldsymbol{\mu}$ y σ).
   b) Enumere los valores de la variable de todas las posibles muestras de tamaño **n=2** y calcule $\bar{\mathbf{x}}$ para cada una de ellas.
   c) Muestre que la media de las 15 medias muestrales ($\bar{\mathbf{x}}$) es $\boldsymbol{\mu}$**.**
   d) Muestre que el desvío estándar de las 15 medias muestrales es:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

   e) ¿Qué puede decirse, en general, sobre el valor anterior cuando el tamaño de la población **N** es muy grande en comparación con el tamaño muestral **n**?

24. Estimación por intervalos

Al comienzo del capítulo del teorema central del límite (TCL) nos preguntamos: ¿Cuál es el peso medio de los recién nacidos? y ¿Cuál es la proporción de alumnos que no entienden estadística? Son muchas más las preguntas que podríamos responder si conociéramos los parámetros de diferentes poblaciones. En la práctica, lo único que podemos hacer es estimarlos; pero las estimaciones tienen error.

¿Entonces?

En vez de **estimar** un parámetro poblacional **mediante** un único número, podemos construir **un intervalo con** una **"garantía"** de cubrir al valor a estimar.

En este capítulo daremos **estimaciones por intervalos** de:

- la media de una población,
- la diferencia de medias de dos poblaciones,
- una proporción poblacional,
- la diferencia de dos proporciones poblacionales.

En todos los casos utilizaremos el TCL para construirlos.

□ 24.1 Intervalos de confianza para la media

Si una variable se distribuye en la población con media μ y desvío estándar σ, se realizan muestreos aleatorios simples y se registran los valores (x_1, ... x_n,) de dicha variable, entonces por el TCL la distribución de muestreo de su **media muestral** ($\bar{x}$) es aproximadamente Normal, con media μ y desvío estándar $\frac{\sigma}{\sqrt{n}}$.

Cerca del 95% de las veces que calculemos una media muestral $\bar{x}$ ésta se encontrará dentro del intervalo

$$\left[\mu - 2\frac{\sigma}{\sqrt{n}};\ \mu + 2\frac{\sigma}{\sqrt{n}}\right]$$

(fig24.1). O sea $\bar{x}$ cumplirá las siguientes desigualdades:

$$\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 2\frac{\sigma}{\sqrt{n}}$$

24.1. *En el eje horizontal se representan los valores de $\bar{x}$. La distribución aproximada es Normal. Cerca del 95% de los valores de muestreo de $\bar{x}$ se encuentran en el intervalo* $\left[\mu - 2\frac{\sigma}{\sqrt{n}};\ \mu + 2\frac{\sigma}{\sqrt{n}}\right]$

Restando miembro a miembro $\bar{x}$+**μ**, y multiplicando miembro a miembro por -1 en las desigualdades anteriores, obtenemos un intervalo para **μ** cuyos límites dependen de $\bar{x}$, **n** y σ:

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2\frac{\sigma}{\sqrt{n}}$$

Este intervalo está **centrado en** $\bar{x}$ y tiene una semiamplitud de **2 desvíos estándar de la media muestral** ($2\frac{\sigma}{\sqrt{n}}$):

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} \qquad \bar{x} \qquad \bar{x} + 2\frac{\sigma}{\sqrt{n}}$$

Cerca del 95% de las veces el intervalo $\left[\bar{x} - 2\frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right]$ **contiene a la verdadera media poblacional μ.**

¿A qué "veces" se refiere la frase anterior? A todas las veces que se realice un muestreo aleatorio simple y se calcule el intervalo.

El 95% de las veces el intervalo contiene a la verdadera media poblacional y 5% no. **Confiamos** en que el intervalo que construimos a partir de una única muestra efectivamente **contenga a μ,** pero no lo podemos saber con certeza. Decimos que:

$IC(\mu) = \left[\bar{x} - 2\frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right]$ es un intervalo de confianza para **μ** del 95% aproximadamente.

El intervalo puede escribirse en forma compacta observando que el mismo término, $2\frac{\sigma}{\sqrt{n}}$, suma o resta $\bar{x}$.

La **media muestral** es un **estimador puntual** de la media poblacional **μ,** es decir que obtenemos un número como estimación y no podemos decir nada respecto a si la media se parece o no se parece a **μ.** Un **intervalo de confianza** también es un **estimador de μ**, pero esta vez tenemos un rango de valores posibles y un grado de confianza (el % de veces que al obtener el intervalo este contendrá a **μ**). La figura 24.2 representa la construcción de intervalos de confianza de la forma $\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$

para la media de una población muchas veces. Conocemos **μ** en este ejemplo y podemos saber cuáles son los intervalos de confianza que contienen a **μ** y cuáles no.

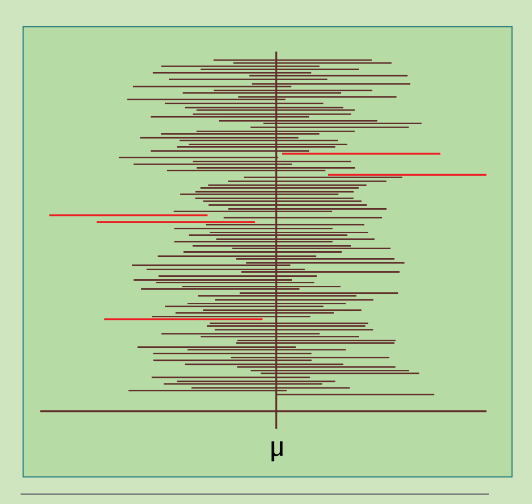


Figura 24.2. Cien intervalos de confianza del 95% para µ, obtenidos de la misma población. Con rojo aparecen los intervalos que no contienen a µ.

Cinco de los 100 intervalos no contienen el valor verdadero del parámetro poblacional que estamos estimando mediante el intervalo.

Un intervalo de confianza, calculado a partir de los datos de una muestra, es uno de los tantos que podríamos obtener a partir de diferentes muestras. Podemos imaginar la construcción de un intervalo de confianza del 95% para µ como una selección al azar de un intervalo, entre todos los intervalos posibles (la figura 24.2 muestra 100 ellos). Algunos intervalos contienen a **µ** son los "buenos".

> ! Los intervalos de confianza no están centrados en **µ; **están centrados en **$\bar{x}$.**

> **Confiamos que el intervalo elegido sea uno de los "buenos"** y no uno "malo" porque estos son sólo el 5%. Pero, **no podemos saber si µ** pertenece al intervalo particular que construimos.

Si queremos aumentar el **nivel de confianza** a 99,7%, aumentamos la longitud del intervalo de confianza tomando 3 desvíos estándar:

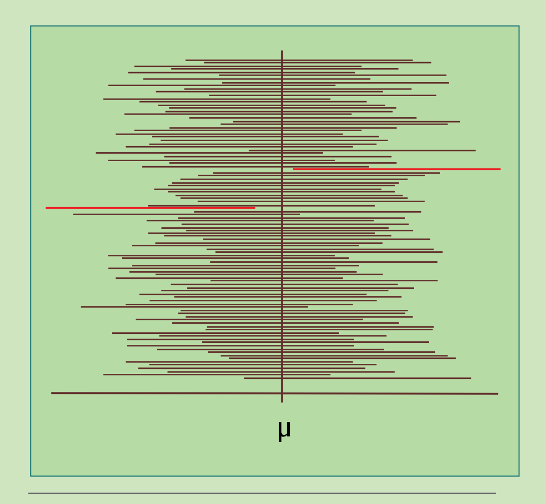


Figura 24.3. Cien intervalos de confianza del 99,7 %. Con rojo aparecen los intervalos que no contienen a µ. Esperamos que a la larga el 99,7% de los intervalos contengan a µ.

> $$\left[\bar{x} - 3\frac{\sigma}{\sqrt{n}}\,;\ \bar{x} + 3\frac{\sigma}{\sqrt{n}}\right]$$
>
> es un **IC (µ) del 99,7%** aproximadamente.

Los intervalos de la figura 24.3 fueron construidos con los mismos datos que los de la figura 24.2 pero tomando $\bar{x} \pm 3\frac{\sigma}{\sqrt{n}}$

de manera que la longitud se multiplicó por un factor 3/2. **Perdemos en precisión** pero **ganamos en confianza.** Dos de los cien intervalos no contienen a **µ.** Si tomáramos más y más intervalos, en el largo plazo, tendríamos un 99,7% de intervalos "buenos".

Ejemplo 3: Supongamos que la fábrica que produce garrafas de gas comprimido de uso doméstico con 20 kg de capacidad nominal está realizando un nuevo proceso de producción, el anterior tenía una media de 47 dm^3. Interesa saber si la media del nuevo proceso ha cambiado, para ello obtiene las capacidades en dm^3 de 50 garrafas:

45,93 47,08 46,32 45,87 46,74 46,44 46,84 45,57 47,09 45,34 46,17 46,45 45,96
47,69 47,51 46,92 46,72 46,39 45,96 46,05 45,94 46,97 46,14 46,79 47,08 47,46
46,45 46,39 47,37 45,98 45,19 45,50 46,14 47,73 46,48 46,71 46,03 45,74 45,99
47,24 45,40 46,56 45,54 46,46 45,44 46,71 46,44 46,67 46,91 45,27

Por las características del proceso se sabe que **σ=0,75** no ha cambiado. Interesa estimar la capacidad media de las garrafas mediante un intervalo de confianza del 95%.

La media muestral de los 50 datos es $\bar{\mathbf{x}}$ = 46,3952 dm^3. Con una confianza del el 95% aproximadamente, podemos decir que la media de las capacidades de todas las garrafas producidas con el mismo proceso se encuentra dentro del intervalo

$$\left[\bar{x} - 2\frac{\sigma}{\sqrt{n}};\ \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right] = [46{,}3952 - 2 \text{ x } 0{,}75 / \sqrt{50}\ ;\ 46{,}3952 + 2 \text{ x } 0{,}75 / \sqrt{50}]$$

$$= [46{,}3952 - 2 \text{ x } 0{,}1061;\ 46{,}3952 + 2 \text{ x } 0{,}1061]$$
$$= [46{,}3952 - 0{,}2122;\ 46{,}3952 + 0{,}2122]$$
$$= [46{,}183;\ 46{,}607]$$

El valor 47 dm^3 no pertenece al intervalo. Si la media de la capacidad en dm^3 no cambió, el 95% de las veces que obtuviéramos un intervalo con otras muestras, pero con el mismo cálculo, el intervalo contendría el valor 47. Sospechamos que el nuevo proceso produce garrafas con menor capacidad.

¿Y si σ es desconocida?

Rara vez el desvío estándar de la distribución de una variable en la población, σ, es conocido. Se utiliza **s** para estimar σ y el **error estándar** ($\frac{s}{\sqrt{n}}$, sección 23.2) para estimar $\frac{\sigma}{\sqrt{n}}$ obteniendo:

$$\left[\bar{x} - 2\frac{s}{\sqrt{n}};\ \bar{x} + 2\frac{s}{\sqrt{n}}\right]$$ un IC (**μ**) **del 95%** aproximadamente.

y $$\left[\bar{x} - 3\frac{s}{\sqrt{n}};\ \bar{x} + 3\frac{s}{\sqrt{n}}\right]$$ un IC (**μ**) **del 99,7%** aproximadamente.

Podemos escribir los intervalos para **μ** de aproximadamente del **95% y 99,7%** nivel de confianza, respectivamente, en forma más compacta:

$\bar{\mathbf{x}}$**±2**.(error estándar) y $\bar{\mathbf{x}}$**±3**.(error estándar).

Nos interesa que la longitud de los intervalos de confianza sea pequeña. Si sumamos y restamos a la media muestral un término muy grande, significa que $\bar{x}$ es una estimación poco precisa de la media poblacional **μ.**

Tres factores afectan la longitud del intervalo de confianza:

- **El nivel de confianza.** Para 95% el desvío estándar de la media se multiplica por 2, y para 99% por 3.
- **El tamaño de la muestra n.** Aparece como $\frac{1}{\sqrt{n}}$. Si n =9, $\frac{1}{\sqrt{n}}$ =1/3=0,333.
- Si n=36, $\frac{1}{\sqrt{n}}$ =1/6=0,166. Para reducir la longitud a la mitad, el tamaño de la muestra debe multiplicarse por 4.
- **La variabilidad de la variable en la población (σ).** A mayor variabilidad de los datos individuales en la población tendremos una longitud mayor del intervalo de confianza.

Ejemplo 3: Continuación.

La encargada del control de procesos en la fábrica de garrafas necesita estimar el volumen de las garrafas con mayor precisión. Desea saber qué tamaño de muestra debe elegir para obtener un intervalo de longitud 0,3 dm^3 con una confianza del 95%.

Sabemos que el intervalo es **$x \pm 2 \frac{\sigma}{\sqrt{n}}$** y **σ=0,75.** Por lo tanto, la longitud del intervalo es:

$$L = 4 \frac{\sigma}{\sqrt{n}}$$

Pero interesa que L=0,3 dm^3, o sea:

$$L = 4 \frac{\sigma}{\sqrt{n}}$$

$$L = 0,3$$

Por lo tanto:

$$n = \left(4 \frac{0,75}{0,3}\right)^2$$

$$n = 100$$

Originalmente se había elegido una muestra de tamaño **n=50,** para obtener la precisión deseada se **debe duplicar ese tamaño de la muestra.**

Podemos preguntarnos si cuánto más pequeña sea la longitud del intervalo, siempre es mejor. Para obtener intervalos de confianza cada vez más estrechos es necesario aumentar más y más el tamaño de la muestra. Esto encarece y dificulta el estudio. En cada situación, el investigador deberá decidir el tamaño de la muestra equilibrando la longitud tolerable para sus intervalos de confianza y sus posibilidades. Estas consideraciones son válidas para todos los intervalos de confianza que presentaremos en el capítulo.

□ 24.2. Intervalos de confianza para la diferencia de medias

El objetivo de muchas encuestas y estudios es comparar dos poblaciones, como los hombres frente a mujeres, los pacientes tratados con la droga A con los tratados con la droga B, las familias de bajos ingresos con las familias de ingresos altos, los individuos con estudios secundarios completos y con estudios secundarios incompletos, respecto de algunas variables. Cuando estas son numéricas (por ejemplo, altura, peso, nivel de colesterol en sangre o ingresos) los parámetros de interés suelen ser sus medias en cada población. Sus estimadores son medias muestrales.

Nuevamente, podemos utilizar el teorema central del límite para hallar la distribución de muestreo de la diferencia de medias muestrales.

Supongamos que una variable numérica se distribuye en una población con media $\boldsymbol{\mu_X}$ y desvío estándar σ_X, y en otra población con media $\boldsymbol{\mu_Y}$ y desvío estándar σ_Y. Si se realizan muestreos aleatorios simples de tamaño $\mathbf{n_X}$ y $\mathbf{n_Y}$ de cada una de las poblaciones respectivamente, se registran los valores (**x**) e (**y**) de dicha variable en cada una de las poblaciones y se calcula la diferencia ($\bar{\mathbf{x}}$-□) de las correspondientes medias muestrales entonces:

- La distribución de muestreo de $\bar{x}-\bar{y}$, es **aproximadamente Normal** con tal de tomar tamaños de muestra n_X y n_Y suficientemente grandes.
- Cuanto mayor sean los tamaños de las muestras ($\mathbf{n_X}$ y $\mathbf{n_Y}$) tanto mejor será la aproximación. Si $\mathbf{n_X}$ y $\mathbf{n_Y}$ son por lo menos 30, la aproximación será buena en la mayoría de los casos.
- La media de la distribución de muestreo de $\bar{x}-\bar{y}$, es $\mu_X-\mu_Y$.
- El desvío estándar de la distribución de muestreo de $\bar{x}-\bar{y}$, es $\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$. Decrece cuando aumentan los tamaños de las muestras ($\mathbf{n_X}$ y $\mathbf{n_Y}$).
- La distribución de muestreo de $\frac{\bar{x}-\bar{y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$ es **aproximadamente Normal** Estándar con tal de tomar tamaños de muestra $\mathbf{n_X}$ y $\mathbf{n_Y}$ suficientemente grandes.

Observación: Nuevamente, **la distribución de la variable no es necesariamente Normal para ninguna** de las dos poblaciones estudiadas. La **distribución** de muestreo de la diferencia de las medias muestrales **sí lo es** aproximadamente, cuando los tamaños de las muestras son suficientemente grandes.

Frecuentemente σ_X y σ_Y no se conocen, entonces se los estima respectivamente, por los desvíos estándar de cada una de las muestras (sección 18.2.3):

$$s_X = \sqrt{\frac{1}{n_X - 1} \sum_{i=1}^{n_X} (x_i - \bar{x})^2} \quad \textbf{y} \quad s_y = \sqrt{\frac{1}{n_Y - 1} \sum_{i=1}^{n_Y} (y_i - \bar{y})^2}$$

El desvío estándar estimado de la distribución de **$\bar{x}$-**□ es:

El error estándar de $\bar{x} - \bar{y} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$

Una estimación para la diferencia de dos medias poblacionales se obtiene tomando la diferencia de medias muestrales (una de cada una de las poblaciones) y sumándole y restándole un margen de error en forma similar a lo que vimos en la sección 24.1 para una única media.

El **intervalo** de nivel **de confianza** aproximado de 95 % **para la diferencia de medias (μ_x-μ_y)** tiene la misma estructura que para una media:

$$IC\ (\mu_X - \mu_Y):\ \bar{x} - \bar{y} \pm 2\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

Ejemplo 4: Llamamos X=peso de un varón de 16 años, Y=peso de una mujer de 16 años. Interesa hallar un intervalo de confianza del 95% para la diferencia de medias de los pesos de varones y mujeres de esa edad.

A partir de los datos del ejemplo 16.4. (supondremos que se trata de pesos provenientes de **muestras representativas** de las poblaciones de varones y mujeres de 16 años de una gran ciudad) obtenemos los siguientes resultados:

Variable	n	Media muestral	Desvío estándar muestral (s)	Error estándar $s/\sqrt{n}$
X	52	66,096	6,6488	0,9220
Y	49	51,265	6,2141	0,8877

Calculemos primero el error estándar de la diferencia de medias

$$\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} = \sqrt{(0{,}9220)^2 + (0{,}8877)^2}$$

$$= \sqrt{1{,}6381}$$

$$= 1{,}28$$

Por lo tanto, un intervalo de confianza del 95% para la diferencia de medias de los pesos de varones y mujeres de 16 años es:

$$\bar{x}-\bar{y} \pm 2\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}$$

$$66{,}1-51{,}3 \pm 2 \,.\, 1{,}28$$
$$14{,}8 \pm 2{,}56$$

Estimamos con una confianza aproximada del 95% que la diferencia de pesos de varones y mujeres de 16 años se encuentra en el intervalo (12,24; 17,36).

□ 24.3. Intervalos de confianza para una proporción

Para cualquier población con una proporción **p** de éxitos y una muestra de tamaño **n** siempre que $np>10$ y $n(1-p)>10$, si $\hat{p} = \frac{\text{cantidad de éxitos en la muestra}}{\text{tamaño de la muestra}}$, el TCL asegura que:

- La distribución de $\hat{p}$ es **aproximadamente Normal.**
- La **media** de la distribución de $\hat{p}$ es **p.**
- El **desvío estándar** de la distribución de **p** es $\sqrt{\frac{p(1-p)}{n}}$ **.**

Por lo tanto, por las propiedades de la curva Normal (sección 20.2.3),

- Cerca del 95% de las proporciones muestrales están entre **p** menos 2 veces el desvío estándar y **p** más 2 veces el desvío estándar, o sea dentro de $p \pm 2\sqrt{\frac{p(1-p)}{n}}$.

- El 99,7% (casi todas) de las proporciones muestrales se encuentran en el intervalo $\left(p-3\sqrt{\frac{p(1-p)}{n}}\,,\ p+3\sqrt{\frac{p(1-p)}{n}}\right)$, o sea dentro de $p \pm 3\sqrt{\frac{p(1-p)}{n}}$.

Podemos realizar cálculos similares a los realizados para la media muestral, pero esta vez para obtener un **intervalo con nivel de confianza de aproximadamente 95% para p:**

Sabemos que cerca del 95% de las veces que calculemos $\hat{p}$

éste se encontrará dentro del intervalo $\left[p-2\sqrt{\frac{p(1-p)}{n}}\ ;\ p+2\sqrt{\frac{p(1-p)}{n}}\right]$

O sea, $\hat{p}$ cumplirá las siguientes desigualdades:

$$p-2\sqrt{\frac{p(1-p)}{n}} \leq \hat{p} \leq p+2\sqrt{\frac{p(1-p)}{n}}$$

Restando miembro a miembro $\hat{p}$**-p** y multiplicando miembro a miembro por -1 en las desigualdades anteriores, obtenemos un intervalo para **p** cuyos límites dependen de $\hat{p}$ y **n**:

$$\hat{p}-2\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p}+2\sqrt{\frac{p(1-p)}{n}}$$

Este intervalo está centrado en $\hat{p}$ y tiene una semiamplitud de **2 desvíos estándar** $2\sqrt{\frac{p(1-p)}{n}}$

$$\hat{p} - 2\sqrt{\frac{p(1-p)}{n}} \qquad \hat{p} \qquad \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}$$

Cerca del 95% de las veces el intervalo $\left[\hat{p} - 2\sqrt{\frac{p(1-p)}{n}} ; \hat{p} + 2\sqrt{\frac{p(1-p)}{n}}\right]$ **contiene a la verdadera proporción poblacional p.**

El intervalo también puede escribirse como $\hat{p} \pm 2\sqrt{\frac{p(1-p)}{n}}$. Pero este intervalo no sirve en la práctica porque no conocemos **p.**

¿Qué se puede hacer? Veamos 2 opciones:

Procedimiento 1. Estimar el desvío estándar $\left(\sqrt{\frac{p(1-p)}{n}}\right)$ por el error estándar $\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ obteniendo el intervalo:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Procedimiento 2. Como $\sqrt{p(1-p)} \leq 0{,}5$ (figura 24.4) podemos acotar el desvío estándar eligiendo el valor más grande de $\sqrt{p(1-p)}\ (0{,}5)$. De esta manera podemos decir: el desvío estándar $\leq 0{,}5 \frac{1}{\sqrt{n}}$. Si tomamos error estándar $= 0{,}5 \frac{1}{\sqrt{n}}$, resulta el intervalo: $\hat{p} \pm \frac{1}{\sqrt{n}}$

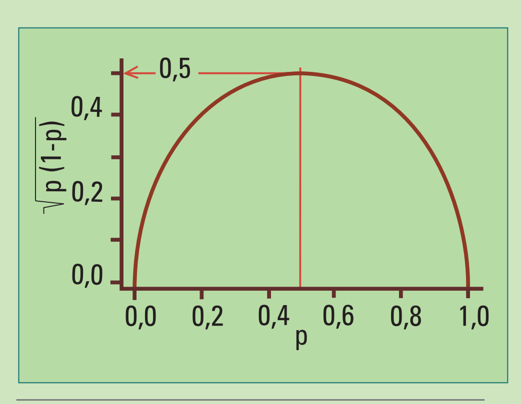


Figura 24.4. Gráfico de la función $f(p)=\sqrt{p(1-p)}$ *para* $0 \leq p \leq 1$.

Se denomina **margen de error** de $\hat{p}$, a la semiamplitud del intervalo de confianza del 95%.

Del procedimiento 1 resulta que el margen de error (sobreentendiendo que se trata del 95%) de $\hat{p}$ es:

$$2 \times 0{,}5 \sqrt{\frac{1}{n}} = \sqrt{\frac{1}{n}}$$

cualquiera sea **p.**

Obtenemos así el

método rápido para el cálculo del **margen de error** de $\hat{p}$ $\qquad \sqrt{\frac{1}{n}}$

Ejemplo 5. Retomemos nuevamente el ejemplo del Club Grande de Fútbol (capítulo 9) la investigadora tomó una muestra de 538 socios y obtuvo $\hat{\mathbf{p}}$**=0,51.** Utilicemos ambos procedimientos para calcular el margen de error:

1. Reemplazar **p** por $\hat{\mathbf{p}}$: $2\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} = 2\sqrt{\frac{0{,}51\ (1-0{,}51)}{538}}$

$$= 2\sqrt{\frac{0{,}2499}{538}}$$

$$= 0{,}04310$$

2. Método rápido. El margen de error $= \sqrt{\frac{1}{538}}$

$$= 0{,}04311$$

Obtuvimos prácticamente el mismo resultado con los dos procedimientos, esto se debe a que $\hat{\mathbf{p}}$**=0,51** está muy cerca de 0,5 donde se alcanza el máximo de $\sqrt{p(1-p)}$ para valores de **p** entre 0 y 1 (figura 24.2). Cuando **p** es pequeño $\frac{1}{\sqrt{n}}$ sobreestima el desvío estándar.

El intervalo de confianza del 95% para la proporción de socios a favor de Rolando Forzudo es: 0,51±0,04, cualquiera sea el procedimiento empleado. El intervalo puede expresarse en porcentajes: 51 %±4 %. Decimos que el resultado de la encuesta es 51% a favor de Rolando Forzudo con un margen de error del 4%. **Confiamos** que el verdadero porcentaje se encuentre dentro del intervalo [50,6; 51,4]. **¿Por qué confiamos?** Porque el 95% de las veces que utilicemos este procedimiento para obtener un intervalo de confianza para una proporción, el mismo contendrá al valor verdadero, pero **no podemos estar seguros** que eso ocurrió **esta vez.**

> Cualquier intervalo de confianza para una proporción puede expresarse en porcentajes, simplemente hay que multiplicar los extremos del intervalo por 100.

> El margen de error del 95% obtenido con el método rápido, $\sqrt{\frac{1}{n}}$ es siempre **mayor o igual** al margen de error calculado con el valor de $\hat{\mathbf{p}}$: $2\sqrt{\frac{\hat{p}\,(1-p)}{n}}$

□ 24.4. Intervalos de confianza para la diferencia de proporciones

Muchos estudios requieren comparar proporciones entre dos poblaciones. Por ejemplo, para comparar la proporción de mujeres y hombres a favor de establecer un salario de desempleo, o la proporción de argentinos que prefiere los automóviles chicos en comparación

con los españoles, o la proporción de pacientes tratados con la droga A que reduce su dolor de cabeza en comparación con los tratados con la droga B, etc.

En todos los ejemplos anteriores tenemos **dos poblaciones** (hombres-mujeres; argentinos-españoles; pacientes tratados con la droga A-pacientes tratados con la droga B) y una característica de interés que llamamos **"éxito".** Éxito puede representar estar a favor un salario de desempleo, preferir automóviles chicos, reducir su dolor de cabeza.

Sean:
$\mathbf{p_1}$ = proporción de éxitos en la población 1
$\mathbf{p_2}$ = proporción de éxitos en la población 2
$\mathbf{n_1}$= tamaño de la muestra de la población 1
$\mathbf{n_2}$= tamaño de la muestra de la población 2
$\hat{\mathbf{p}}_1$ = proporción de éxitos en la muestra de la población 1
$\hat{\mathbf{p}}_2$ = proporción de éxitos en la muestra de la población 2

También se puede aplicar el TCL para la distribución de muestreo de la diferencia de proporciones:

- Si $\mathbf{n_1\, p_1}$, $\mathbf{n_1\,(1\text{-}p_1)}$, $\mathbf{n_2\, p_2}$, $\mathbf{n_2\,(1\text{-}p_2)}$ son todos mayores a 10, entonces la distribución de muestreo de $\hat{p}_1 - \hat{p}_2$ es **aproximadamente Normal.**
- La **media** de la distribución de $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$ es $\mathbf{p_1 - p_2}$.
- El desvío estándar de la distribución de $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$ es $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
- La distribución de muestreo de $\dfrac{\hat{p}_1 - \hat{p}_1 - (p_1 - p_1)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ es **aproximadamente Normal Estándar** si $\mathbf{n_1\, p_1}$, $\mathbf{n_1\,(1\text{-}p_1)}$, $\mathbf{n_2\, p_2}$, $\mathbf{n_2\,(1\text{-}p_2)}$ son todos mayores a 10.

En la práctica no se conocen $\mathbf{p_1}$ ni $\mathbf{p_2}$, por lo cual la condición de los tamaños muestrales ($\mathbf{n_1\, p_1}$, $\mathbf{n_1\,(1\text{-}p_1)}$, $\mathbf{n_2\, p_2}$, $\mathbf{n_2\,(1\text{-}p_2)}$ todos mayores a 10) se verifica reemplazando $\hat{\mathbf{p}}_1$ y $\hat{\mathbf{p}}_2$ por **y.**

! **Las muestras deben:**
- ser independientes entre sí,
- ser a lo sumo un 10% de la población,
- provenir de un muestreo aleatorio simple.

Luego el intervalo de aproximadamente el 95% de confianza, para la diferencia de proporciones ($\mathbf{p_1\text{-}p_2}$) basado en las proporciones muestrales, es:

$$IC\,(p_1\text{-}p_2): \hat{p}_1 - \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

y el intervalo es aproximadamente el 99,7%, para la diferencia de proporciones (p_1 -p_2) basado en las proporciones muestrales es:

$$IC(p_1-p_2): \hat{p}_1-\hat{p}_2 \pm 3\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

La diferencia de proporciones muestrales ($\hat{\mathbf{p}}_1$- $\hat{\mathbf{p}}_2$) es un **estimador** de la diferencia de las proporciones poblacionales $\mathbf{p_1}$-$\mathbf{p_2}$ cuyo **desvío estándar** es:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Ejemplo 6: Supongamos que en un muestreo aleatorio simple de 125 choferes de taxis de la Ciudad de Buenos Aires, el 64% opina que el riesgo de sufrir una agresión de parte de un pasajero aumentó durante el último año. Mientras que de una encuesta similar realizada entre 150 choferes de taxis del Conurbano Bonaerense, el 76% dio esa misma respuesta. Construya un intervalo de confianza del 95% para la verdadera diferencia de proporciones.

Sean:
Población 1: todos los choferes de taxis de la Ciudad de Buenos Aires.
Población 2: todos los choferes de taxis del Conurbano Bonaerense.
"Éxito" = opinar que el riesgo de sufrir una agresión de parte de un pasajero aumentó durante el último año
$\mathbf{p_1}$ = proporción de éxitos en la población 1
$\mathbf{p_2}$ = proporción de éxitos en la población 2
$\mathbf{n_1}$= tamaño de la muestra de la población 1 =125
$\mathbf{n_2}$= tamaño de la muestra de la población 2 =150
$\hat{\mathbf{p}}_1$ = proporción de éxitos en la muestra de la población 1 = 0,64
$\hat{\mathbf{p}}_2$ = proporción de éxitos en la muestra de la población 2 = 0,76

Primero:

$n_1\hat{p}_1 = (125)(0,64)$ $\quad$ $n_2\hat{p}_2 = (150)(0,76)$

$n_1\hat{p}_1 = 80$ $\quad$ $n_2\hat{p}_2 = 114$

$n_1(1-\hat{p}) = (125)(1-0,64)$ $\quad$ $n_2(1-\hat{p}) = (150)(0,76)$

$n_1(1-\hat{p}) = 45$ $\quad$ $n_2(1-\hat{p}) = 36$

Son todos mayores a 10.

Segundo: Las muestras son menos del 10% de la población de choferes de taxis, tanto para la Ciudad de Buenos Aires como para el Conurbano Bonaerense.

Tercero:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0,64(1-0,64)}{125}+\frac{0,76(1-0,76)}{150}}$$

$$= 0,0553$$

$$\hat{p}_1 - \hat{p}_2 = 0,64 - 0,76$$
$$= -0,12$$

¡La diferencia de proporciones puede ser negativa!

Finalmente, el intervalo resulta

$$\hat{p}_1 - \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}}{n_2}}$$

$$-0,12 \pm 2 \times 0,0553$$

$$-0,12 \pm 0,1106$$

$$(-0,2306 ; -0,0094)$$

Podemos asegurar con un 95% de confianza que la diferencia de proporciones de los choferes de taxis que opinan que el riesgo de sufrir una agresión de parte de un pasajero aumentó durante el último año, se encuentra entre -0,2306 y -0,0094. Con un 95% de confianza podemos decir que la proporción es menor en la Ciudad de Buenos Aires.

Ejemplo 7: Retomemos el ejemplo 10 de la sección 22.5. (Dos variables Categóricas) en el que interesa estudiar si existe asociación entre dos variables categóricas: “come rápido” y “sobrepeso” ambas con categorías “sí, “no”. Dentro del grupo de individuos que come rápido el 30 % tiene sobrepeso, mientras que entre los que no comen rápido ese porcentaje se reduce al 10 %. Los porcentajes de individuos con sobrepeso en las dos categorías de la variable come rápido son bastante diferentes. Necesitamos saber si esa diferencia puede atribuirse a la variabilidad que surge del muestreo para decidir que las variables están asociadas.

Sean:
Población 1: todos los individuos que comen rápido.
Población 2: todos los individuos que no comen rápido.
“Éxito” = tener sobrepeso.
$\mathbf{p_1}$ = proporción de éxitos en la población 1
$\mathbf{p_2}$ = proporción de éxitos en la población 2
$\mathbf{n_1}$= tamaño de la muestra de la población 1 = 250
$\mathbf{n_2}$= tamaño de la muestra de la población 2 = 220
$\hat{\mathbf{p}}_1$ = proporción de éxitos en la muestra de la población 1 = 0,3
$\hat{\mathbf{p}}_2$ = proporción de éxitos en la muestra de la población 2 = 0,1

Primero:

$n_1 \hat{p}_1 = (250)(0,30)$ $\quad$ $n_2 \hat{p}_2 = (220)(0,10)$

$n_1 \hat{p}_1 = 75$ $\quad$ $n_2 \hat{p}_2 = 22$

$n_1 (1-\hat{p}_1) = (250)(0,70)$ $\quad$ $n_2 (1-\hat{p}_2) = (220)(0,90)$

$n_1 (1-\hat{p}_1) = 175$ $\quad$ $n_2 (1-\hat{p}_2) = 198$

son todos mayores a 10.

Segundo: Las muestras son menos del 10% de las poblaciones en consideración

Tercero:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-p)}{n_2}} = \sqrt{\frac{0{,}30(1-0{,}30)}{250}+\frac{0{,}10(1-0{,}10)}{220}}$$

$$= 0{,}0939$$

$$\hat{p}_1 - \hat{p}_2 = 0{,}30 - 0{,}10$$

$$\hat{p}_1 - \hat{p}_2 = 0{,}20$$

Finalmente, el intervalo resulta: $\hat{p}_1 - \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$$0{,}20 \pm 2 \times 0{,}0939$$

$$0{,}20 \pm 0{,}1878$$

$$(0{,}0122\ ;\ 0{,}3878)$$

Como el cero no es un valor del intervalo, podemos asegurar con un 95% de confianza que las proporciones poblacionales $\mathbf{p_1}$ **y** $\mathbf{p_2}$ son distintas. Esto es que las variables "sobrepeso" y "come rápido" están asociadas.

□ 24.5. Consideraciones generales sobre intervalos de confianza

Todos los intervalos de confianza presentados tienen **la misma forma:**

Estimador ± K x desvío estándar del estimador.

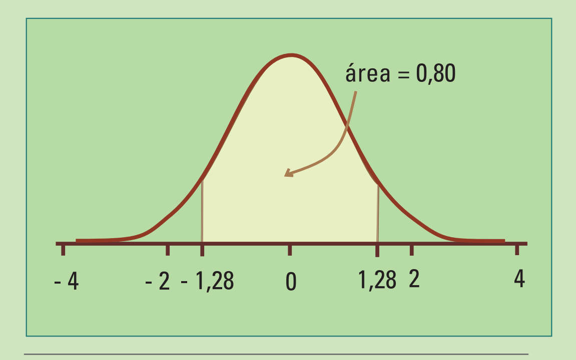
Hemos utilizado la regla 68-95-99,7 para obtener intervalos con niveles aproximados de confianza del 95% y 99,7% tomando **K=2** y **K=3** respectivamente. Pero, puede ocurrir que consideremos que un 95% de confianza es un criterio demasiado exigente para nuestro problema en particular, o que necesitamos un nivel de confianza mayor. Es posible obtener intervalos con **cualquier nivel de confianza** entre 0 y 100% utilizando diferentes valores de **K.** La tabla 24.1 presenta diferentes valores de **K,** y sus correspondientes niveles de confianza.

En particular, la figura 24.3 muestra el área de 80% bajo la curva Normal entre -1,29 y 1,29; si se elige K=1,29 para calcular el intervalo de confianza para un parámetro utilizando la forma:

Estimador ± K x desvío estándar del estimador.

NIVELES DE CONFIANZA Y VALORES DE K OBTENIDOS DE LA DISTRIBUCIÓN N(0,1). TABLA 24.1

Niveles de confianza	K	Niveles de confianza	K
50 %	0,67	95,000%	1,96
60 %	0,84	99,000%	2,58
70 %	1,04	99,000%	3,29
80 %	1,28	99,990%	3,89
90 %	1,64	99,999%	4,42



24.5. *Área bajo la curva Normal entre -1,28 y 1,28.*

La regla 68-95-99,7 utiliza el 2 como aproximación al valor **K=1,96.** En la mayoría de las aplicaciones esta aproximación es adecuada, pero la tabla 23.2 muestra el valor más exacto 1,96.

Ejemplo 7. Continuación.
Construyamos el intervalo del 95% de confianza para la diferencia de porcentajes de individuos con sobrepeso en las dos categorías de la variable come rápido, pero esta vez utilizando 1,96 en vez de 2:

$$0{,}20 \pm 1{,}96 \times 0{,}0939$$
$$0{,}20 \pm 0{,}1840$$
$$(0{,}016;\ 0{,}384)$$

El resultado es esencialmente el mismo que habíamos obtenido antes.

24.6. Actividades y ejercicios

1. Una fábrica produce dardos con diámetros que tienen un desvío estándar σ= 0,25mm. De un lote grande, se seleccionó una muestra aleatoria simple de 40 dardos, y se obtuvo un promedio de 3,09 mm en sus diámetros.

a. Obtenga un intervalo de confianza de aproximadamente el 95% para la media de los diámetros de todos los dardos producidos de la misma forma.
b. Idem b pero con una confianza aproximada del 99.7%
c. De qué tamaño debe ser la muestra para que la longitud del intervalo de confianza del 95% sea 0,1 mm.

En los ejercicios 2 a 7 se presentan varias respuestas, elija la correcta y explique brevemente.

2. Cambiar de 95% a 99,7% el nivel de confianza de un intervalo para una proporción, dejando el resto igual.

a. Aumenta la longitud del intervalo en 4,7%
b. Reduce la longitud del intervalo en 4,7%
c. Aumenta la longitud del intervalo en 50%
d. Reduce la longitud del intervalo en 50%
e. No puede saberse sin conocer el tamaño de la muestra.

3. Se prueban 49 autos de un nuevo modelo y se registran los litros de nafta consumidos en un recorrido de 100 km, obteniéndose una media muestral, **x=6,8** litros y un desvío estándar muestral, **s=1,4** litros. Obtenga un intervalo de aproximadamente 95% de confianza para la cantidad media de litros de nafta consumida por ese tipo de vehículo en 100 km.

a. [5,4; 8,2]
b. [6,6; 7,0]
c. [6,4; 7,2]
d. [6,2; 7,4]

4. Se sabe que el 82 % de los alumnos del último año de las escuelas secundarias dependientes de alguna universidad planean seguir estudios superiores. Supongamos que se selecciona una muestra aleatoria simple de alumnos del último año de dichas escuelas y se obtiene un intervalo de confianza en base a la proporción que manifiesta tener interés en continuar sus estudios. Entonces:

a. El centro del intervalo de confianza es 0,82.
b. El intervalo de confianza contiene el valor 0,82.
c. Un intervalo de confianza del 99,7% contiene el valor 0,82.

5. En general, ¿cómo cambia la longitud del intervalo de confianza si se duplica el tamaño de la muestra y todo lo demás queda igual?

a. Se duplica la longitud.
b. La longitud se reduce a la mitad
c. La longitud se multiplica por 1,414
d. La longitud se divide por 1,414
e. No se puede saber.

6. Una encuesta reveló que el porcentaje de mujeres que no le gusta planificar sus vacaciones con más de un mes de anticipación es 68% con un margen de error del ± 5%. ¿Qué significa el ± 5%?

a. Se encuestó al 5 % de la población.
b. En la muestra, el porcentaje de mujeres que no le gusta planificar sus vacaciones con más de un mes anticipación se encontró entre 63% y 73%.
c. En la población, el porcentaje de mujeres que no le gusta planificar sus vacaciones más de un mes con anticipación está entre 63% y 73%.
d. Se encuestó entre 63% y 73% de la población.
e. Sería raro que en la población el porcentaje de mujeres que no le gusta planificar sus vacaciones con más de un mes anticipación esté fuera del intervalo de 63% a 73%.

7. En un muestreo aleatorio simple de 300 hombres mayores de 70 años, el 48 % eran viudos y en un muestreo aleatorio simple de 400 mujeres en el mismo rango de edades, 65% eran viudas. Halle un intervalo de confianza para la diferencia entre el porcentaje de viudas y viudos.

a) 17% ± 0,38 %
b) 17% ± 7,48 %
c) 55,6 % ± 7,48 %
d) 56,5 % ± 0,74 %
e) 56,5 % ± 0,38 %

8. Se realizó un muestreo aleatorio simple, de un embarque de 50.000 piezas delicadas, registrándose 16 piezas dañadas de un total de 220 observadas. Obtenga un intervalo del 95% de confianza para estimar la verdadera proporción y a partir de él la cantidad de piezas dañadas.

9. El desvío estándar de la distribución de muestreo de $\hat{\mathbf{p}}$ es $\sqrt{\frac{p(1-p)}{n}}$, depende de **p**. Para entender cómo se comporta para distintos valores de **p**, grafique en el eje vertical $\sqrt{p(1-p)}$ y en el eje horizontal **p** para los siguientes valores de **p**: 0 0,1 0,2 0,3 ... 0,9 1. Trace una curva que una los puntos. Observe que el gráfico alcanza su máximo para **p=0,5.** El margen de error calculado con $\hat{\mathbf{p}}$ ¿será menor o igual que el obtenido por el procedimiento rápido?

25. Decisiones en el campo de la estadística

Muchas afirmaciones que escuchamos a diario conciernen al campo de la estadística:

- Las lentejas instantáneas requieren solamente 2 minutos de hervor para estar listas para comer.
- La vida media de las computadoras es 6 años.
- El consumo moderado de alcohol en las comidas reduce el riesgo de infarto de miocardio.
- El 20% de las mujeres maneja mal.

Se trata de afirmaciones respecto de una o varias poblaciones.

¿Qué se puede hacer si alguna de las afirmaciones nos concierne especialmente?

- Creerle o no creerle, directamente.
- Realizar nuestra propia investigación, siguiendo los lineamientos propuestos en este libro.
- Indagar respecto de cómo se llegó a la conclusión.

Creer (o no creer) directamente en los resultados no es una buena opción. Realizar su propia investigación, eso es lo que realizan muchas investigadoras de mercado e institutos médicos, farmacéuticos, universidades. Para la tercera opción, es necesario saber qué hay que mirar para evaluar el estudio y entender los resultados.

Generalmente la afirmación se refiere a un parámetro poblacional; es decir, un número que caracteriza a toda la población. Por ejemplo, la afirmación el "20% de las mujeres maneja mal" se refiere a la población de todas las mujeres que conducen un automóvil particular o algún otro tipo de vehículo. Se está realizando una afirmación sobre un porcentaje (el parámetro) referido a todas las mujeres que conducen un automóvil particular o algún otro tipo de vehículo (la población). ¿Puede alguien saber exactamente cuál es ese porcentaje? Nadie puede saberlo; por lo tanto, esa afirmación no corresponde necesariamente a un hecho real. Se trata de una hipótesis, se la denomina "hipótesis nula" y es necesario validarla. Pero uno puede tener su propia hipótesis basada en su experiencia. Por ejemplo, puede ser: "el porcentaje de mujeres que maneja mal es menor a 20%"; es la "hipótesis alternativa". También se puede cuestionar la hipótesis nula diciendo simplemente: "el porcentaje no es 20%".

Además de realizar "pruebas de hipótesis" sobre una variable categórica (maneja mal: sí, no), se pueden realizar también sobre variables continuas, como la cantidad de años que dura una computadora sin averiarse. En estos casos el parámetro de interés es la media poblacional; la hipótesis nula será una afirmación respecto del valor del parámetro (por ejemplo $\mu=6$ años, en general $\mu=\mu_0$ con μ_0 un número fijo).

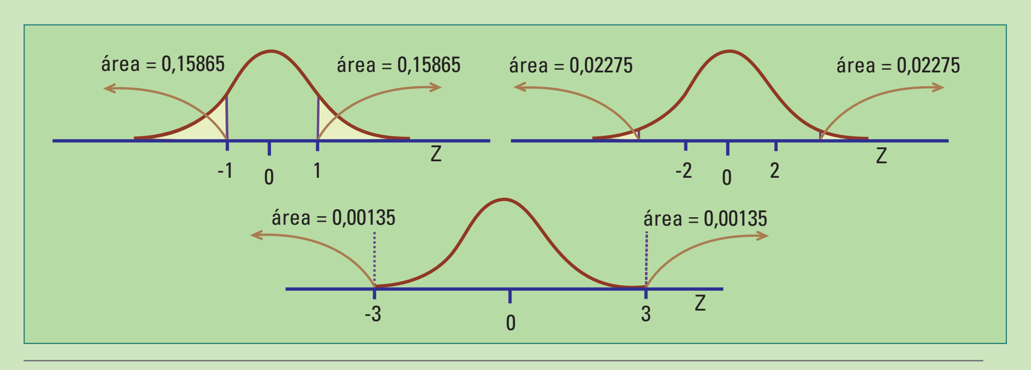
25.1. Prueba de hipótesis

La idea detrás de "la prueba de hipótesis", también conocida como "el test de hipótesis", es pensar que **si la hipótesis nula,** por ejemplo: **$\mu=\mu_0$, fuera verdadera** la diferencia entre la media muestral **($\bar{x}$)** y **μ_0** debería ser pequeña. Si es demasiado grande para lo esperable, debido simplemente a la variabilidad aleatoria, se sospecha de la validez de la hipótesis nula y se toma la decisión a favor de la hipótesis alternativa. Es un razonamiento similar al que se utiliza en demostraciones por el **método de reducción al absurdo;** se supone válido lo contrario de lo que se quiere probar para llegar a algo falso o contradictorio. En el caso del test de hipótesis, los datos proveen la información para sospechar que el supuesto (hipótesis nula) es falso o no lo es.

Para saber si la diferencia entre, por ejemplo, la media muestral ($\bar{x}$) y el valor especificado en la hipótesis nula (μ_0) es grande se necesita **construir un estadístico a** partir de **$\bar{x}-\mu_0$**, cuya **distribución** de muestreo **sea conocida cuando la hipótesis nula es verdadera.** Así podremos decidir si la diferencia lleva a un valor razonablemente posible de la distribución o se trata de un valor raro. Frente a esta última situación podemos sospechar que la hipótesis nula no es verdadera.

Utilizaremos estadísticos cuya distribución es aproximadamente Normal Estándar cuando la hipótesis nula es verdadera. Recordemos que cuando un conjunto de datos se distribuye en forma Normal, la mayor concentración de ellos se encuentra en el centro de la distribución, y a medida que nos alejamos del centro hacia las colas la concentración disminuye más y más.

Al construir intervalos de confianza utilizamos la regla 68-95-99,7, para obtener en forma aproximada las áreas de 3 zonas centrales bajo la curva de densidad Normal. También pueden considerarse las áreas complementarias; hacia las colas de la distribución Normal esperamos encontrar aprox. el 32%, 5% y 0,3% de los datos a medida que nos alejamos 1 desvío estándar, 2 desvíos estándar y 3 desvíos estándar de la media. Más precisamente se trata de 15,865%, 2,275% y 0,135% del área en cada cola de la distribución Normal Estándar (figura 25.1).



25.1. *Las áreas de las colas, bajo la curva Normal Estándar, disminuyen a medida que nos alejamos del cero.*

Encontrar el valor observado del estadístico en una zona de la distribución donde esperamos muy pocos datos, cuando la hipótesis nula es verdadera, es una evidencia en su contra.

Ejemplo 1. Nos dicen que la vida media de las computadoras es 5 años pero sospechamos que es menor.

Consideramos la afirmación: "la vida media de las computadoras es 5 años" como la **hipótesis nula** (**H_0**, se lee hache cero) y a nuestra sospecha: "la vida media de las computadora es menor a 5 años" como **hipótesis alternativa** (**Ha**).

Llamando **µ** a la media de la vida de todas las computadoras, entonces las hipótesis se escriben:

$$\mathbf{H_0: \mu = 5} \text{ y } \mathrm{Ha}: \mu < 5$$

Una vez que se establecen la hipótesis nula y la alternativa, el paso siguiente consiste en hallar la evidencia para tomar la decisión. La calidad de los datos es fundamental; la información debe ser precisa y no tener sesgo. Una mayor precisión se obtiene con un mayor tamaño de muestra: para evitar el sesgo los datos deben provenir de un muestreo aleatorio simple.

Supongamos que elegimos **n=36** computadoras al azar obteniendo una media muestral $\bar{x}$ = 4,33 años y un desvío estándar **s=1,12 años.** Queremos decidir si la diferencia entre 4,33 y el valor especificado en la hipótesis nula (5) es atribuible al azar (debido a haber tomado una muestra), o tenemos suficiente evidencia para rechazar la hipótesis nula a favor de la alternativa planteada.

Debemos construir el **estadístico del test. Si la hipótesis nula es verdadera** y como n> 30 sabemos que la media muestral $\bar{x}$ tiene una distribución aproximadamente Normal con media **µ=5** y desvío $\frac{s}{\sqrt{n}} = \frac{1,12}{6}$ (sección 23.1).

Por lo tanto el **estadístico del test** es: $z = \frac{6(\bar{x} - 5)}{s}$ y tiene una distribución (de muestreo) aproximadamente Normal Estándar (**N**(0,1)) si **H_0: µ=5** es verdadera.

El **valor del estadístico del test** es:

$$z_{observado} = \frac{6(4,33 - 5)}{1,12}$$

$$= -3,5892$$

¿Provee el valor -3,5892 del estadístico del test suficiente evidencia para rechazar la hipótesis nula a favor de la alternativa?

Si la hipótesis nula fuera cierta -3,5892 sería un valor observado de una Normal Estándar. Pero **el área por debajo de la Normal Estándar para valores menores o iguales a -3,5892** es 0,0001658.

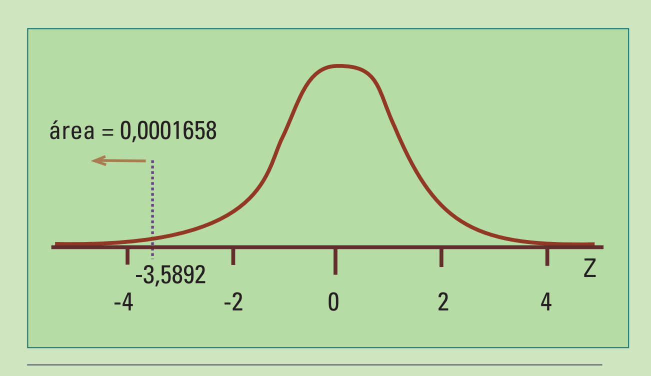


Figura 25.2. *El área por debajo de la curva N(0,1) para valores más extremos que -3,5892 es 0,0001658.*

Esto significa que si la hipótesis nula fuera cierta, obtendríamos menos de dos veces cada 10.000 un valor tan o más extremo como -3,5892, **en la dirección de la hipótesis alternativa** -hacia los valores menores-. ¡Es una frecuencia muy baja! Decimos que la diferencia entre la media muestral y el valor nulo (**μ=5**) no es atribuible al azar. Por lo tanto, sospechamos de la validez de la hipótesis nula.

Conclusión. Rechazamos la hipótesis nula y decimos que los datos proveen suficiente evidencia a favor de la hipótesis alternativa: **Ha: μ<5.** La vida media es "significativamente" menor que 5 años.

En este contexto, "significativamente" expresa que el resultado se obtuvo a partir de un test de hipótesis y que la diferencia entre el valor especificado en la hipótesis nula (5) y el observado (4,33) no es atribuible al azar. No se está diciendo nada respecto a la magnitud de la diferencia.

□ 25.2. Valor-p

El área, por debajo de la Normal Estándar para valores tan o más extremos como **el valor observado del estadístico del test** en dirección de la hipótesis alternativa, se llama **valor-p.** Corresponde a la **proporción de valores** del estadístico del test, tan o más extremos, que se obtendrían como resultado del muestreo aleatorio si la hipótesis nula fuera verdadera. **Cuanto más pequeño sea el valor-p, tanto mayor será la evidencia a favor de la hipótesis alternativa.**

En el ejemplo 1 el **valor-p** $(z_{observado} = -3{,}5892) = 0{,}0001658$ (ver figura 25.2).

> En la práctica los tests de hipótesis se realizan con computadoras utilizando **programas estadísticos** que **calculan los valores-p.**

Ejemplo 2. Un estudio nacional obtiene que la presión sistólica media es 129, para varones entre 35 y 44 años de edad. Un médico laboral de una aseguradora de trabajo sospecha que los ejecutivos tienen una presión sistólica elevada debido al estrés laboral. Selecciona al azar los registros de 45 ejecutivos en ese grupo etáreo y obtiene una media muestral **$\bar{x}$=130,8** y un desvío estándar **s=17,2.** ¿Tiene el médico suficiente evidencia para decir que la media de la presión de los ejecutivos es mayor que la de la población en general?

Las **hipótesis** son:

H_0: μ=129; la media de la presión sistólica de todos los ejecutivos entre 35 y 44 años no difiere de la media nacional de los varones del mismo grupo etáreo.
Ha: μ>129; la media de la presión sistólica de todos los ejecutivos entre 35 y 44 años es mayor a la media nacional de los varones del mismo grupo etáreo.

Se plantea una alternativa unilateral en dirección a los valores mayores **(μ>129)** porque el médico considera que el estrés únicamente puede aumentar o dejar igual la presión sistólica, nunca reducirla.

Como **n>30** y los datos provienen de un muestreo aleatorio simple, **el estadístico del test** es:
$z = \frac{\sqrt{n}\,(\bar{x} - 129)}{s}$ y tiene distribución aproximadamente **N(0,1)** si **H_0: μ=129** es verdadera (sección 24.1)

El valor observado del estadístico del test es:

$$z_{observado} = \frac{\sqrt{45}\,(130{,}8 - 129)}{17{,}2} = 0{,}7020$$

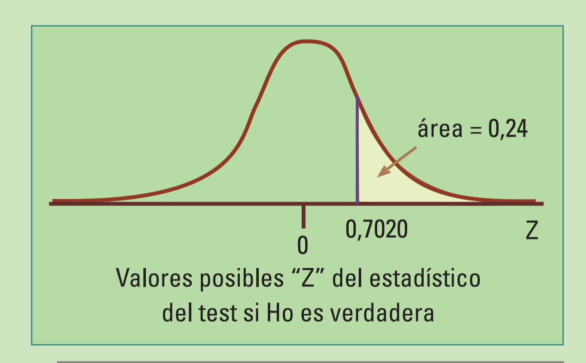


Figura 25.3. *Curva de densidad N(0,1) y $z_{observado}$ =0,7020. El área bajo la curva por encima de ese valor es 0,24.*

El valor-p, para $z_{observado}$=0,7020, es 0,24 (figura 25.3); o sea, el área bajo la curva de densidad Normal Estándar en la dirección de la hipótesis alternativa es 0,24. Aproximadamente 25% de los valores de la distribución de muestreo del estadístico del test estarían por encima del valor observado si **H_0** fuera verdadera; la diferencia entre el valor especificado en la hipótesis nula y la media muestral no es tan grande, es atribuible a la variabilidad debida al muestreo aleatorio.

Conclusión. No hay razones para sospechar que la media de la presión sistólica de todos los ejecutivos entre 35 y 44 años es mayor a la media nacional de los varones del mismo grupo etáreo.

Ejemplo 4. Retomemos el ejemplo 10 de la sección 22.5. (Dos variables Categóricas) en el que interesa estudiar si existe asociación entre dos variables categóricas: "come rápido" y "sobrepeso" ambas con categorías "sí, "no". Entre 250 individuos que "come rápido" se encuentra una proporción muestral de $\hat{p}_1$=0,3 individuos con sobrepeso y entre 220 individuos que "no come rápido" $\hat{p}_2$=0,1 Queremos decidir si la diferencia de proporciones observada de individuos con sobrepeso entre las categorías "come rápido" y "no come rápido" es atribuible al azar.

Llamamos Población 1 a todos los individuos que comen rápido y Población 2 a todos los individuos que no comen rápido y "éxito" es "tener sobrepeso".

Las **hipótesis** son:

H_0: p_1-p_2 = 0; Las proporciones poblacionales de "tener sobrepeso" son iguales entre los que comen rápido y los que no comen rápido.
Ha: p_1- p_2 ≠ 0; Las proporciones poblacionales de "tener sobrepeso" son diferentes entre los que comen rápido y los que no comen rápido.

donde $\mathbf{p_1}$ = la proporción de éxitos en la población 1
$\mathbf{p_2}$ = la proporción de éxitos en la población 2

> **Muchas veces** los tests de hipótesis se utilizan para comparar parámetros de dos poblaciones y el **valor especificado en la hipótesis nula es cero,** de allí su nombre.

> Al **valor especificado en la hipótesis nula** se lo denomina **valor nulo, sea o no sea cero.**

La hipótesis alternativa es bilateral. Se plantea este tipo de alternativa porque el efecto de comer rápido, en principio podría estar en ambos sentidos.

Como **n_1 p_1, n_1 (1-p_1), n_2 p_2, n_2 (1-p_2)** son todos mayores a 10 la distribución de muestreo de **$\hat{p}_1$-$\hat{p}_2$** es aproximadamente Normal (sección 24.4).

Por lo tanto **el estadístico del test** es:

$$\frac{\hat{p}_1 - \hat{p}_1 - 0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

y tiene una distribución de muestreo **aproximadamente Normal Estándar** cuando **H_0: p_1-p_2 = 0** es verdadera

El **valor observado del estadístico del test es:**

$$z_{observado} = \frac{0{,}30 - 0{,}10}{\sqrt{\dfrac{0{,}30(1-0{,}30)}{250} + \dfrac{0{,}10(1-0{,}10)}{220}}}$$

$$z_{observado} = \frac{0{,}20}{0{,}0939}$$

$$z_{observado} = 2{,}129$$

El valor-p es 0,033. Como se trata de un test bilateral (también llamado a dos colas), el **valor-p** es el área bajo la curva de la distribución de muestreo del estadístico del test para valores mayores a 2,1299, y hacia los valores menores de -2,1299 (figura 25.4).

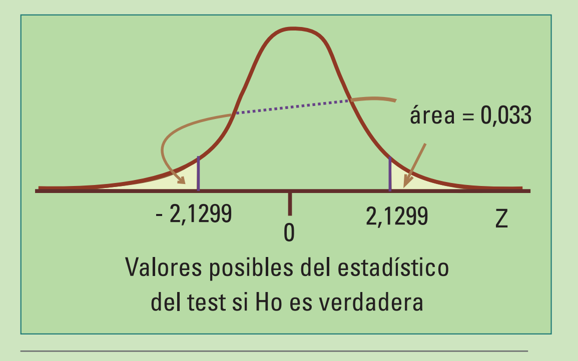


Figura 25.4. *Área de dos colas. El valor-p, para $z_{observado}$=2,1299 es 0,033 y la curva N(0,1)*

El área de las dos colas es 0,033. Aproximadamente 3 de cada 100 veces obtendríamos una diferencia entre el valor observado y el valor nulo tan o más extrema, consideramos que es una frecuencia baja.

Conclusión. Se rechaza la hipótesis de igualdad de proporciones. Decimos que la diferencia de las proporciones muestrales no es atribuible al muestro aleatorio y por lo tanto es estadísticamente significativa.

25.2.1 Áreas por debajo de la curva Normal estándar

Muchos programas estadísticos calculan áreas bajo la curva Normal Estándar. Los cálculos aquí se realizan utilizando el lenguaje **R** (ver recuadro). Para obtener un **área bajo la curva Normal Estándar a la izquierda de un valor cualquiera (z)** se escribe **pnorm (z)** y el programa devuelve el área.

El p-valor del ejemplo 1 (figura 25.2) se obtuvo de la siguiente manera:

Escribiendo **pnorm** (-3,5892) en la consola del programa el resultado devuelto es 0,0001658.

Si interesa el **área a derecha** de **z,** como el total del área bajo una curva de densidad es 1, simplemente se escribe **1-pnorm (z).** En el ejemplo 2 el área por encima del valor 0,7020 es: 1-pnorm (0,7020) = 1- 0,7586604 = 0,2413396. Este número se redondeó a 0,24.

R

Se trata de un lenguaje de programación integrado con muchos programas para manipulación de datos, cálculo y visualización gráfica. Es un entrono en el que se implementan distintas técnicas estadísticas. R (http://www.r-project.org/) es el resultado de un esfuerzo de colaboración con las contribuciones de todo el mundo y es de distribución libre. Inicialmente R fue escrito por Robert Gentleman y Ross Ihaka- también conocidos como "R & R" del Departamento de Estadística de la Universidad de Auckland.

Más importante que saber cómo se calcula es saber qué significa **un valor-p.**

25.3. Nivel de significación

Muchas veces se decide de antemano cuán pequeño debe ser el valor-p para declarar que la diferencia entre el valor especificado en la hipótesis nula y el valor observado es estadísticamente significativo. Ese valor se lo llama nivel de significación, y se lo indica por la letra griega **α (alfa).**

Al elegir **α=0,05** se permite que a lo sumo un 5% de las veces se rechace en forma equivocada la hipótesis nula. Pero si **α=0,01** la exigencia de la prueba es mayor, a lo sumo se rechazará en forma equivocada la hipótesis nula un 1% de las veces.

Si al realizar un test, el valor-p resulta tan pequeño como o menor que α (valor-p ≤ α), decimos que: **la diferencia observada** con el valor especificado en la hipótesis nula **es estadísticamente significativa a nivel α.**

En el ejemplo 4 el p-valor = 0,033. Entonces, la diferencia observada es estadísticamente significativa a nivel **α=0,05** y no lo es a nivel **α=0,01.**

En el ejemplo 1 el p-valor = 0,0001658. Entonces, la diferencia es estadísticamente significativa a nivel **α=0,05** y también a nivel **α=0,01.**

Pero el **p-valor dice mucho más que el nivel de significación,** proporciona el **menor nivel** que con los datos observados, **el test resultaría en rechazo.**

Los niveles de significación habituales son el 5% y el 1%, pero el **valor-p** es más informativo.

□ 25.4. Decisiones en base a intervalos de confianza

Cuando la hipótesis alternativa es bilateral, es posible tomar la decisión de rechazar la hipótesis nula utilizando un intervalo de confianza. Se rechaza la hipótesis nula si el valor especificado en dicha hipótesis (valor nulo) no se encuentra dentro del intervalo de confianza del parámetro o diferencia de parámetros sobre los que se quiere tomar la decisión.

Un **intervalo de confianza** es más informativo que un nivel de significación o un valor-p porque provee además **una estimación** del parámetro o diferencia de parámetros.

Para un test de nivel de significación **α** debe utilizarse un intervalo de nivel de confianza **100 x (1-α)%.** El nivel de significación del test controla la proporción de equivocaciones al rechazar la hipótesis nula, interesa que sea lo mas pequeño posible. El nivel de confianza, en porcentaje, expresa el porcentaje de aciertos en los cuales el intervalo contiene al parámetro verdadero; interesa que sea grande.

Ejemplo 5. De los registros de los últimos 40 meses de dos sucursales de una gran empresa se calcularon las cantidades medias por mes de accidentes **($\bar{x}_1$=8,6; $\bar{x}_2$=7,5)** y sus correspondientes desvíos estándar **(s_1=2,5; s_2=2,4),** para decidir si las medias muestrales difieren significativamente.

Las **hipótesis** son:

H_0: μ_1-μ_2 = 0; Las medias poblacionales de la cantidad de accidentes son iguales en las dos sucursales.
Ha: μ_1-μ_2 ≠ 0; Las medias poblacionales de la cantidad de accidentes de las dos sucursales difieren.

La hipótesis alternativa es bilateral porque no tenemos razones, más allá de los datos, para pensar que la diferencia tenga que estar necesariamente en algún sentido.

Como el tamaño de las muestras es suficientemente grande la distribución de muestreo de $\bar{x}_1$-$\bar{x}_2$ es aproximadamente Normal (sección 24.2). Por lo tanto el **estadístico del test** es: $z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{40} + \frac{s_2^2}{40}}}$ tiene una distribución de muestreo N(0,1) **cuando H_0: μ_1-μ_2 = 0 es verdadera.**

El valor observado del estadístico del test es:

$$z_{observado} = \frac{8,6 - 7,5}{\sqrt{\frac{2,5^2}{40} + \frac{2,4^2}{40}}}$$

$$z_{observado} = \frac{1,1}{0,548}$$

$$z_{observado} = 2,007$$

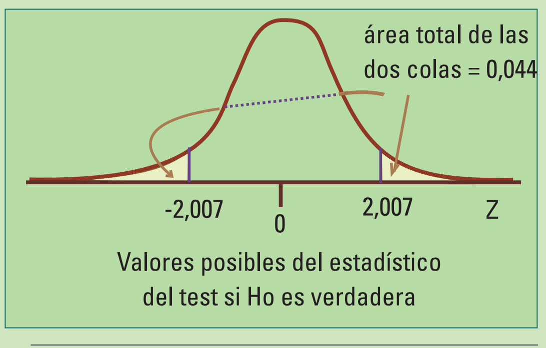


Figura 25.5. *El área bajo la curva Normal Estándar para z<-2,007 y z>2,007.*

valor-p = 0,0446986

Conclusión. Como el valor-p ≤ 0,05 se rechaza la hipótesis de igualdad de medias. Decimos que la diferencia de las medias muestrales no es atribuible al muestro aleatorio y es estadísticamente significativa a nivel **α=0,05.**

Podemos también construir un intervalo de aproximadamente el 95% de confianza para la diferencia de medias poblacionales **(μ_1-μ_2)** utilizando la diferencia de las medias muestrales **($\bar{x}_1$-$\bar{x}_2$):**

$$IC\,(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm 1,96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}}$$

$$= 8,6 \;-7,5 \pm 1,96 \times 0,548$$

$$= 1,1 \pm 1,074$$

$$= [0,026;\ 2,174]$$

El intervalo no contiene al valor nulo (es cero en este ejemplo), por lo tanto la diferencia de accidentes entre las dos sucursales es significativamente diferente a nivel **α=0,05.** Es la misma decisión que obtuvimos al realizar el test calculando su valor-p. Pero ahora tenemos un rango de valores posibles para esa diferencia. Quien realice el estudio además de saber que la diferencia es significativa podrá decidir, en base al un rango de valores posibles para esa diferencia, si la diferencia de la cantidad de accidentes por mes entre las sucursales justifica tomar medidas para eliminarla.

Observación En el cálculo del IC **(μ_1- μ_2)** estamos utilizando el valor 1,96 (tabla 24.1) en vez de 2 de la regla 68-95-99,7.

□ 25.5. Expresiones generales

En las secciones anteriores se presentaron ejemplos de diferentes pruebas de hipótesis. Todas tienen las siguientes componentes:

a) Hipótesis nula - Hipótesis alternativa.
b) Estadístico en base al que se toma la decisión llamado **estadístico del test.**
c) Cálculo del valor-p, área bajo la curva Normal Estándar de valores tan o más extremos como el observado del estadístico del test.
d) Nivel **α.**

La tabla 25.1 muestra la hipótesis nula y los estadísticos de los tests **(z)** obtenidos a partir de muestras de tamaño grande para distintos parámetros poblacionales. Todos los estadísticos tienen distribución aproximadamente N(0,1) cuando la hipótesis nula es verdadera.

Para cualquiera de los tests planteados el cálculo del valor-p (tabla 25.2) depende del tipo de alternativa planteada.

TESTS DE HIPÓTESIS SOBRE PARÁMETROS POBLACIONALES UTILIZANDO TAMAÑOS DE MUESTRAS GRANDES. TABLA 25.1

Tests para	Hipótesis nula (H0)	Estadístico del test
Una media poblacional **(μ)**	$\mu = \mu_0$	$z = \sqrt{n}\,\frac{\bar{x} - \mu_0}{s}$
Una proporción poblacional **(p)**	$p = p_0$	$z = \sqrt{n}\,\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})}}$
La diferencia de medias poblacionales **(μ_1-μ_2)**	$\mu_1 - \mu_2 = 0$	$z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
La diferencia de proporciones poblacionales **(p_1-p_2)**	$p_1 - p_2 = 0$	$z = \frac{\hat{p}_1 - \hat{p}_1 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

Valor-p para las distintas hipótesis alternativas (Ha). TABLA 25.2

Bilateral	Unilateral hacia los valores mayores	Unilateral hacia los valores menores
$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
$p \neq p_0$	$p > p_0$	$p < p_0$
$\mu_1 - \mu_2 \neq 0$	$\mu_1 - \mu_2 > 0$	$\mu_1 - \mu_2 < 0$
$p_1 - p_2 \neq 0$	$p_1 - p_2 > 0$	$p_1 - p_2 < 0$
Valor-p **Área** por debajo de la curva **Normal Estándar** de los valores menores a $-\lvert z \rvert_{observado}$ y mayores a $\lvert z \rvert_{observado}$ valor - p = área $\lvert z \rvert_{observado}$ $\lvert z \rvert_{observado}$	**Valor-p** **Área** por debajo de la curva **Normal Estándar** de los valores mayores a $\lvert z \rvert_{observado}$ valor - p = área $z_{observado}$	**Valor-p** **Área** por debajo de la curva **Normal Estándar** de los valores menores a $\lvert z \rvert_{observado}$ valor - p = área $z_{observado}$

Los intervalos de confianza y las pruebas de hipótesis presentadas utilizan el teorema central del límite para establecer la distribución de muestreo de varios estadísticos. Esa es la razón por la que el cálculo de todos los valores-p presentados en la tabla 24.2 requieren cálculos de áreas bajo la N(0,1).

Sin embargo, existen otros resultados de la teoría de probabilidad que permiten describir la distribución de muestreo de los estadísticos considerados para tamaños de muestras chicas pero incorporando nuevos supuestos. Se trata de nuevas distribuciones a partir de las cuales se obtienen intervalos de confianza y tests de hipótesis.

No sólo es posible obtener intervalos de confianza y realizar pruebas de hipótesis para los parámetros estudiados, sino también para otros parámetros como, por ejemplo, el desvío estándar que no hemos tratado. Pero todos los conceptos desarrollados siguen siendo válidos.

25.6. Actividades y ejercicios

En los ejercicios con varias respuestas elija la que mejor responde a la pregunta planteada, o completa la afirmación.

1. Cuando se rechaza la hipótesis nula con un valor-p = 0,03 eso significa que

   a. La hipótesis nula no es verdadera.
   b. El 3% de las veces que utilice el test la hipótesis nula no será verdadera.
   c. El 97% de las veces que utilice el test la hipótesis nula no será verdadera.
   d. Cuando la hipótesis nula sea verdadera sólo el 3% de las veces se obtendrá un valor tan o más extremo que el observado.

2. La hipótesis alternativa es bilateral cuando

   a. No existen razones para suponer que los resultados necesariamente tendrán una dirección.
   b. Cuando no se realizó un experimento previo para establecer la dirección del efecto que se quiere probar.
   c. Cuando lo que se quiere probar es pequeño.

3. Se plantea una hipótesis alternativa es unilateral porque

   a. Los datos así lo sugieren.
   b. Existen razones intrínsecas al problema estudiado por las cuales el efecto sólo puede ocurrir en una dirección.
   c. Es más fácil calcular el valor-p.
   d. La cantidad de datos no es suficiente para plantear una hipótesis bilateral.

Los ejercicios siguientes que requieran del cálculo de valores-p, utilice los datos de la figura 25.6 para hallarlos en forma aproximada.

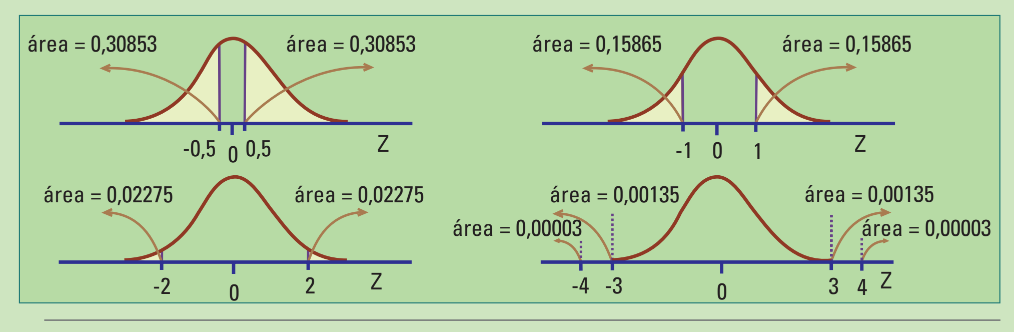


Figura 25.6. Las áreas de las colas de la curva Normal Estándar disminuyen a medida que nos alejamos del cero.

4. El intendente afirma que el tiempo medio que tardan las ambulancias del servicio de emergencias de la ciudad desde que recibe el pedido hasta llegar al lugar del hecho es de 12 minutos. Un periodista sospecha que el tiempo en realidad es mayor, porque se trata de una ciudad muy grande y nunca puede ser menor a 12 minutos. ¿Qué hipótesis nula y qué alternativa debe plantear?

a. H_0: $\mu = 0$ y Ha: $\mu = 12$
b. H_0: $\mu = 12$ y Ha: $\mu > 12$
c. H_0: $\mu = 0$ y Ha: $\mu \neq 0$
d. H_0: $\mu = 12$ y Ha: $\mu < 12$
e. H_0: $\mu = 12$ y Ha: $\mu \neq 12$

5. Ejercicio 4 continuación. El periodista obtiene el tiempo que tardaron las ambulancias en llegar al lugar del accidente utilizando los registros de 44 accidentes,. Si la media muestral obtenida fue de 18 minutos con un desvío de 8 minutos, ¿qué puede decirse de la afirmación del intendente?

a. El valor-p es menor a 0,00006 indicando una evidencia muy fuerte en contra de la afirmación del intendente.
b. El valor-p es 0,02 de manera que la evidencia en contra de lo afirmado por el intendente es significativa a nivel **α=0,05** pero no a nivel **α=0,01.**
c. El valor-p es 0,09 indicando alguna evidencia en contra de lo afirmado por el intendente.
d. El valor-p es 0,49 indicando ninguna evidencia en contra de lo afirmado por el intendente.

6. Se realizó un estudio para determinar si hay diferencias de opinión entre habitantes de ciudades chicas (con menos de 10.000 habitantes) y ciudades grandes (con más de 100.000 habitantes). De una muestra de 140 habitantes seleccionados de varias ciudades chicas se obtuvo que el 68 pensaban que el cambio climático va afectar fuertemente sus vidas en los próximos 20 años mientras que entre los 180 habitantes seleccionados de varias ciudades grandes fueron 95 los que tuvieron esa opinión. ¿Cuál de las siguientes hipótesis son las adecuadas en este estudio?

a. H_0: $p = 68/140$ y Ha: $p = 95/180$
b. H_0: $p_1 - p_2 = 68/140 - 95/180$ y Ha: $p_1 - p_2 < 68/140 \neq 95/180$
c. H_0: $p_1 - p_2 = 0$ y Ha: $p_1 - p_2 \neq 0$
d. H_0: $p_1 - p_2 = 0$ y Ha: $p_1 - p_2 < 0$

7. En relación a los datos del ejercicio 6.

a. El valor del estadístico del test es -2 por lo tanto el valor-p = 0,02275.
b. El valor del estadístico del test es muy cercano a -4 por lo tanto el valor-p es muy cercano a 0,00003 dando altísima evidencia que las proporciones entre ciudades grandes y chicas son diferentes.

c. El valor del estadístico del test es - 0,714. Por lo tanto el valor - p > 0,30.
No hay suficiente evidencia para concluir que las opiniones difieren.

8. Realice una encuesta entre los profesores y alumnos de su escuela para saber si piensan que el cambio climático va afectar fuertemente sus vidas en los próximos 20 años. Compare mediante tests de hipótesis e intervalos de confianza las diferentes proporciones entre alumnos y profesores. Repita la comparación, pero esta vez entre varones y mujeres.

9. Indique si las afirmaciones son verdaderas o falsas.

a. Un valor-p grande muestra una alta evidencia en contra de la hipótesis nula.
b. Un tamaño de muestra grande compensa el sesgo porque se utiliza el Teorema Central del Límite para hallar la distribución del estadístico del test.
c. La hipótesis alternativa depende de los datos.
d. Si se rechaza la hipótesis nula a nivel **α=0,01** entonces también se la rechaza a nivel **α=0,05.**
e. Si se rechaza la hipótesis nula a nivel **α=0,05** entonces también se la rechaza a nivel **α=0,01.**

26. Epílogo: estadística y probabilidad.

La teoría de probabilidad comenzó hacia 1654, cuando el jugador Chevalier de Méré consultó a Blaise Pascal por sus pérdidas inesperadas (ver recuadro Paradoja de Chevalier de Méré). Pascal, junto con su amigo Pierre de Fermat, explicaron las aparentes contradicciones y sentaron las bases de la teoría de probabilidad mediante un intercambio epistolar.

Pero fue recién a partir de la formulación axiomática de la teoría de probabilidad (Kolmogorov, 1933) que pudo definirse probabilidad en términos matemáticos precisos.

Paradoja de Chevalier de Méré

Los jugadores franceses del siglo 17 solían apostar a obtener por lo menos 1 as en cuatro tiradas de un dado. Una modificación de ese el juego consistía en arrojar 2 dados 24 veces y la apuesta era sobre la aparición de por lo menos un doble as.

De acuerdo con el razonamiento de Chevalier de Méré, un as tiene una chance de 1/6 en una tirada (que es correcto) y 4/6 en 4 tiradas (que es incorrecto). Para 2 tiradas razonaba en forma similar: 1 / 36 son las chances de 2 aces en dos tiradas (que es correcto). Luego para compensar deben realizarse 24 tiradas obteniendo nuevamente un resultado incorrecto: 24/36 = 4/6 (el mismo resultado que en el juego con 4 tiradas)

¿Cuáles son los resultados correctos?

Probabilidad de tener por lo menos un as en 4 tiradas de un dado.

Al arrojar un dado equilibrado 4 veces se pueden tener 6 x 6x 6 x 6 = 6^4 resultados equiprobables. De esos 5 x 5 x 5 x 5 = 5^4 no tienen ningún as. Por lo tanto 6^4 - 5^4 son resultados favorables a la apuesta. Luego la probabilidad de obtener por lo menos un as en 4 tiradas es:

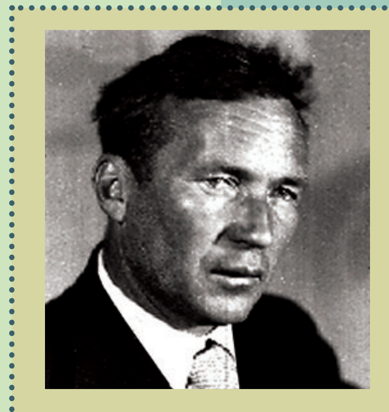
$$(6^4 - 5^4)/6^4 = \frac{(1.296 - 625)}{1.296} = \frac{671}{1.296} = 0{,}51775$$

Esta probabilidad es levemente mayor a 1/2 favoreciendo al apostador.

Probabilidad de tener por lo menos un doble as en 24 tiradas de dos dados.

Tirar dos dados tiene 36 resultados equiprobables, 35 de los cuales son desfavorables en la apuesta. En 24 tiradas hay 36^{24} resultados posibles de los cuales (36^{24} - 35^{24}) son favorables. Por lo tanto, la probabilidad de ganar la apuesta en este caso es:





Andrey Nikolaevich Kolmogorov (1903 -1987). Matemático ruso que realizó, entre muchas otras, importantes contribuciones en estadística y teoría de probabilidad.



$(36^{24} - 35^{24}) / 36^{24} = 1 - (35/36)^{24}$
$\approx 1 - 0{,}5086$
$= 0{,}4914$

Esta probabilidad es levemente menor a 1/2 y por lo tanto desfavorable al apostador.

El enfoque sobre la variabilidad de los datos es lo que distingue la estadística de la matemática. Pero, ¿Qué es lo que distingue la teoría de probabilidad de la estadística, siendo que ambas estudian fenómenos aleatorios? Veamos dos situaciones:

- **Situación 1.** Se sabe que la mitad de los socios del "Club Grande de Fútbol" apoyan al candidato Rolando Forzudo. Se eligen 5 socios al azar ¿Cuántos socios estarán a favor de ese candidato?
- **Situación 2.** Se pregunta ¿Qué proporción de los socios del "Club Grande de Fútbol" apoyan al candidato Rolando Forzudo?

Ninguna respuesta es determinista. La elección de un socio al azar produce un resultado aleatorio en ambas situaciones.

En la situación 1 se comienza suponiendo que la mitad de los socios apoya un candidato. A partir de allí **se realiza una deducción lógica utilizando un modelo** teórico, para obtener las probabilidades de todos los resultados posibles, 0, 1, 2, ... etc.

En la segunda situación se comienza **sin saber** qué proporción de socios realmente está a favor de Rolando Forzudo. Para hallar una respuesta se selecciona una muestra aleatoria simple y se estima la proporción de socios a favor del candidato a partir de la misma. En este caso **la respuesta se induce a partir de observaciones experimentales.**

Desarrollar un modelo para un experimento aleatorio, como por ejemplo para el experimento de arrojar una moneda n veces, y obtener fórmulas para el cálculo de las probabilidades de los distintos resultados (Modelo Binomial) forma parte de la **teoría de probabilidad.** El modelo suele tener parámetros; construir un método para estimar esos parámetros -utilizando datos-, hallar sus propiedades y determinar bajo qué condiciones son válidas se encuentran dentro del campo de la **estadística teórica.** Aplicar el método, verificando las condiciones de validez del mismo, forma parte de la **estadística aplicada.**

El Modelo Binomial es muy general, depende de un único parámetro (**p** = probabilidad de cara) y de la cantidad de veces que se arroja la moneda **(n).** Arrojar una moneda y observar si salió cara o ceca es equivalente a elegir un individuo al azar de una población con dos categorías (éxito, fracaso) y observar a qué categoría pertenece. Por ejemplo, en el caso de los socios del Club Grande de Fútbol (Éxito= socio que apoya la candidatura de Rolando Forzudo, Fracaso= socio que no apoya la candidatura).

Es parte de la estadística teórica demostrar que $\hat{\mathbf{p}}$ es un buen estimador de **p** y hallar sus propiedades. Quien utilice ese estimador, antes de sacar sus conclusiones,

deberá verificar que lo está haciendo en las condiciones correctas. Los datos deben provenir de un muestreo aleatorio simple sin sesgo y la variable que se está observando debe ser una medida válida de la característica en estudio. Si el tamaño de la muestra es más del 10% del tamaño de la población además deberá utilizar un factor de corrección.

A medida que se profundiza en el estudio de estadística son cada vez más necesarios los conocimientos de matemática y probabilidad.

El diseño, la recolección de datos, así como su análisis y la interpretación de los resultados, son aspectos fundamentales de la estadística. Dependen fuertemente del contexto y -en niveles introductorios- requieren de poco uso de matemática formal. Esto permitió presentar el pensamiento estadístico, y su aplicación a la resolución de problemas, explicando y cuantificando la variabilidad de los datos sin utilizar explícitamente cálculos de probabilidad.

Cuando las preguntas se complican las respuestas requieren de herramientas estadísticas más complejas. Hay un mundo de posibilidades para quienes se atrevan. Esto es sólo el comienzo.

27. Respuestas y soluciones

Capítulo 3

3.1. Si tomamos como referencia el PBI por habitante, la relación se reduce a que el PBI por habitante de Brasil es el 84% del de Argentina.

3.2. La cantidad de conductores es generalmente mayor entre las 18 h y 20 h (hora pico) que entre las 14 h y las 16 h. Es esperable que la cantidad de accidentes sea mayor. La proporción de accidentes es una medida más adecuada que la cantidad en este caso.

3.3. Se trata de noticias como:

- Los argentinos comen muy mal: exceso de carne y poca verdura. Se come un 75% más de carnes rojas de lo recomendado y un 46% menos de verduras.
- La cena navideña es 10 veces más calórica de lo sugerido.
- Referido al fútbol. Entre los cinco grandes suman 87 campeonatos, el 76% de los títulos.

3.4. Se trata de avisos como los de los ejercicios 5 y 7.

3.5. ¿Qué significa que la piel naranja sea menos visible? ¿Cómo se mide? ¿El producto fue efectivo en todas las mujeres? ¿Cómo se mide que una piel esté lisa? ¿Cómo se obtuvo el -1,9cm? ¿Es un promedio? ¿Es la reducción mínima? ¿Es la reducción máxima? ¿Las 44 que realizaron la autoevaluación están entre las 50 que participaron del test clínico?

3.6. Explique las siguientes frases:

- Le puedo pagar a lo sumo $500 por ese trabajo: Le puede pagar $500 pero no más de ese valor
- Le voy a pagar como mínimo $500 por ese trabajo. Le puede pagar $500 y tal vez más.
- Quiero que vuelvas como máximo a las 11 de la noche. A las 11 o antes de esa hora.
- Se presentaron por lo menos 10 personas para el puesto de encargado de control de calidad. Se presentaron 10, 11, 12, etc. pero no se sabe cuántos exactamente.
- No más de 10 personas se presentaron para el puesto de chofer. Pueden haberse presentado: ninguna, una, dos, ..., ó 10 personas pero no más.

3.7. “Un chico de 8 a 12 años puede perder hasta un litro de transpiración durante dos horas de actividad un día caluroso”, afirma una publicidad.

- Un procedimiento para estimar cuanto líquido puede perder un chico por transpiración durante dos horas:

Pese varios chicos antes y después de realizar una actividad física durante 2 horas.

- “Hasta dos litros” significa que puede
- no perder nada
- perder 1 litro
- perder 2 litros
- perder 1,5 litros

Capítulo 6

6.1.

- Una encuesta de opinión contacta a 1.243 adultos y les pregunta, ¿ha comprado un billete de lotería en los últimos 12 meses? **Población:** todos los adultos. No aclara ninguna característica. **Muestra:** 1.243 adultos.
- Durante la reunión anual del colegio de abogados, todos los presentes (2.500), llenaron una encuesta referida al tipo de seguro que prefería para su automóvil. **Población:** 2.500 miembros del colegio de abogados.
- En 1968 se realizó en Holanda un test de inteligencia a todos los varones de 18 años que estaban realizando el Servicio Militar Obligatorio. **Población:** todos los varones de 18 años de Holanda en 1968.
- El INDEC lleva a cabo la Encuesta Permanente de Hogares (EPH) en la que se encuestan 25.000 hogares para captar información sobre la realidad económico-social de la República Argentina. **Población:** todos los hogares. **Muestra:** 25.000 hogares.

6.2.

- Sesgos por subcubrimiento. Podrían, por ejemplo, quedar sub-representados sectores sociales con mayores dificultades para movilizarse o aquellos que piensan que van a perder ya que tendrán menor entusiasmo para realizan el esfuerzo de ir a votar.
- Sesgo para tratar de agradar o para no quedar mal con los demás miembros del bloque.

6.3.

Indique cuál es el tipo de muestreo realizado en cada caso.

- Cada alumno escribe su nombre en un papel, lo pone en una bolsa y el director elige 100 papeles. Se trataría de un muestreo aleatorio simple, pero con este procedimiento surge la pregunta de cuán bien fueron mezclados los nombres en la bolsa. Podrían salir elegidos más alumnos de un curso porque los papeles estarían juntos.
- A cada alumno se le asigna un número entre 1 y 2.500 y se seleccionan generando 100 números al azar de cuatro dígitos utilizando algún programa de computación: muestreo aleatorio simple.
- Para cada año, se asigna a cada alumno un número entre 1 y 500, y se elige 1 de cada 25 alumnos: muestreo sistemático.
- Se eligen al azar una división de cada uno de los años y se seleccionan 20 alumnos de cada división: muestreo en dos etapas.
- Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años: muestreo estratificado.

- Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años. Se seleccionan en forma separada los varones y las mujeres de acuerdo con la proporción de mujeres y varones que tiene la escuela: muestreo proporcional al género.

6.4.

a) Que la encuesta no dice nada porque arrastra el sesgo por respuesta voluntaria.

6.5.

a) VÁLIDA
b) NO VÁLIDA
c) VÁLIDA
d) NO VÁLIDA
e) VÁLIDA
f) NO VÁLIDA

□ Capítulo 7

7.1.

a. Ambas dimensiones de la figura, alto y ancho, fueron reducidas a la mitad, el área es 0,5 x 0,5 = 0,25 veces más pequeña. Nuestros ojos responden al área que se redujo a la cuarta parte. Para que el área se reduzca a la mitad la altura y el ancho deben corregirse por el factor $\sqrt{0,5} = 0,707$. Es importante que la relación de los valores se muestre mediante áreas; son éstas las que realmente producen la impresión visual.

b. La relación entre los valores de los bancos está magnificada, produciendo una sensación de mayor caída que la real. El artista representó los valores de los bancos por los diámetros de los círculos. Por ejemplo, para el BNP Paribas la relación entre los diámetros es

$$\frac{D_2}{D_1} = \frac{108}{32,5} = 3,32$$

(D_1 = diámetro del círculo menor y D_2 = diámetro del círculo mayor). Los radios mantienen la misma relación,

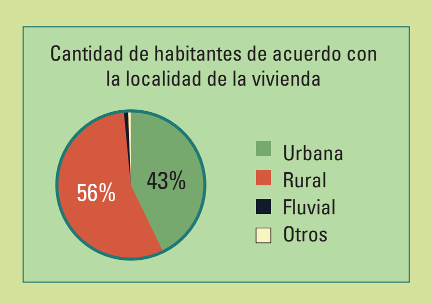
$$\frac{R_2}{R_1} = 3,32$$

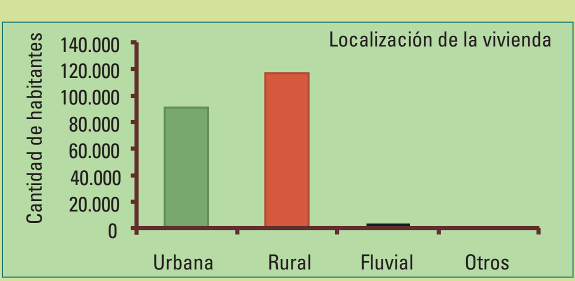
El cociente entre las respectivas áreas es

$$\frac{\text{Área1}}{\text{Área2}} = \left(\frac{R_2}{R_1}\right)^2 = 3,32^2 = 11,02$$

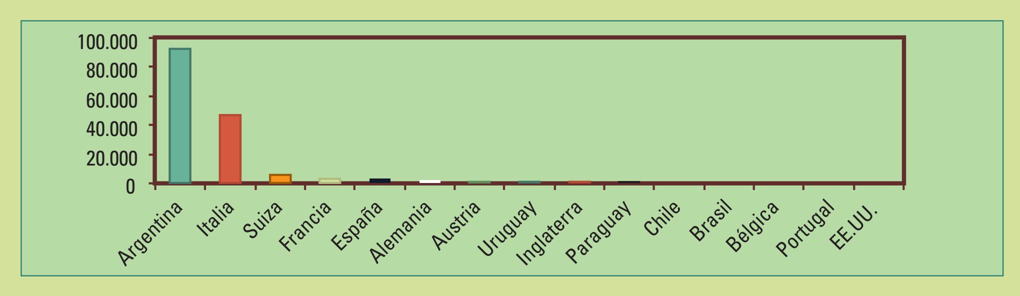
Visualmente el año 2007 aparece como 11 veces mayor, cuando en realidad lo es un poco más de 3 veces.

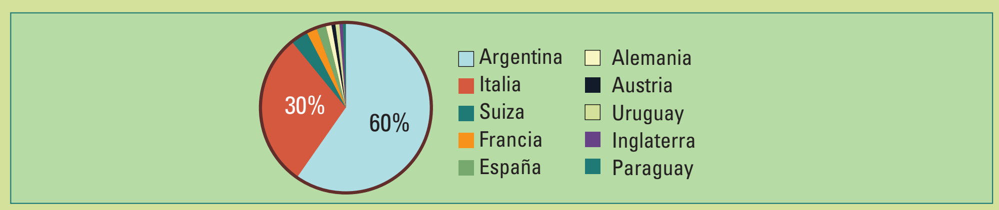
7.2. Los valores de las tres variables consideradas (Alfabetización, Nacionalidades y Localización) pueden representarse tanto en un gráfico de barras, como en un diagrama circular porque se trata de partes de un total.





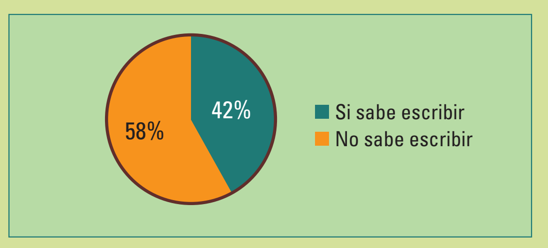
El 99% de la población es Urbana o Rural





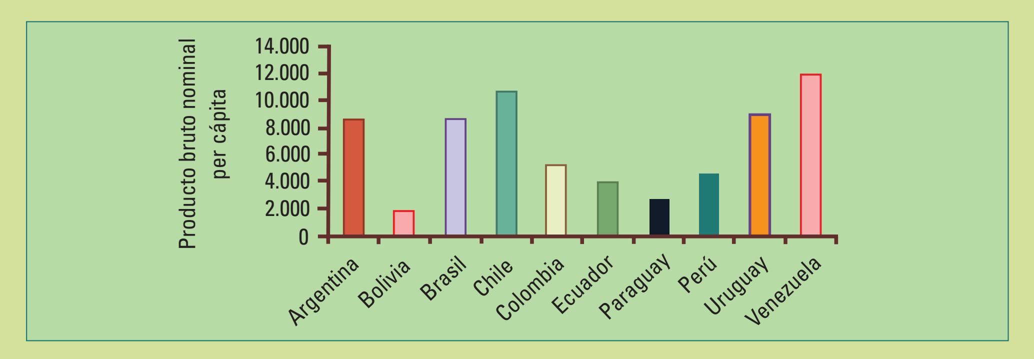
Dos nacionalidades abarcan al 90% de la población.





En 1887, más de la mitad de la población de la Provincia de Santa Fe no sabía escribir.

7.3. La altura de las barras representa el producto bruto total de cada país dividido por la cantidad de habitantes.



No se puede reemplazar la figura anterior por un gráfico circular porque los datos no representan una parte de un total.

Capítulo 8

8.1.

- ¿Están contentos los alumnos con el nuevo sistema de promoción? Una encuesta, porque interesa conocer la opinión de los alumnos sin modificarla.
- ¿El ausentismo de los alumnos es menor en verano que en invierno? Estudio observacional que no es una encuesta, se trata de describir un comportamiento.
- ¿El rendimiento de los alumnos en un examen es mejor si durante el mismo escuchan música de Vivaldi, en bajo volumen, en comparación con no escuchar nada? Estudio experimental, porque interesa comparar los resultados de aplicar dos "tratamientos": 1) con música de Vivaldi, 2) sin música de Vivaldi.

8.2. Similares a los del ejercicio 1.

8.3. Es un experimento porque está eligiendo a los niños y niñas al azar para dividirlos en los dos grupos a los que les enseñará utilizando canciones y a los que no. Ni los niños (ni sus padres) eligen en qué grupo participar.

Capítulo 9

9. 1.
Unidad muestral: una arandela
Variable: continua, diámetro de una arandela
Tamaño de la Muestra: 100

Población: todas las arandelas del lote
Parámetro: 1,908 cm
Estadístico: media muestral
Valor del estadístico: 1,915 cm

9. 2.
Unidad muestral: familia, el enunciado no especifica ni el año ni la región.
Variable: categórica con dos categorías: 1) madre sabe que un antibiótico no puede curar un resfrío, 2) madre no sabe que un antibiótico no puede curar un resfrío
Tamaño de la Muestra: 213
Población: todas las familias
Estadístico: porcentaje de madres que saben que un antibiótico no puede curar un resfrío
Valor del estadístico: 40%

9. 3.
Unidad muestral: hogar de la Argentina en el año 2001
Variable: categórica con dos categorías: 1) hogar heladera con freezer, 2) hogar sin heladera con freezer
Población: todos los hogares de la Argentina en el año 2001
Parámetro: porcentaje de hogares con heladera con freezer
Valor del parámetro: 50%

9. 4.
Unidad muestral: un auto
Variable: continua, el precio de un auto
Tamaño de la Muestra: 8
Estadístico: media muestral
Valor del estadístico: $ 21.880

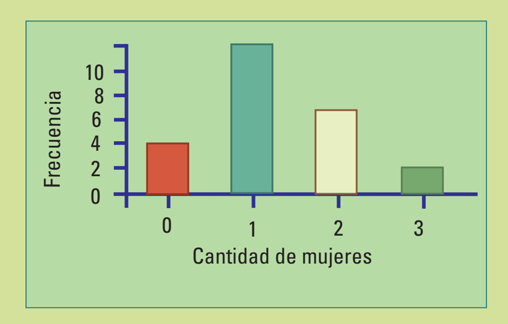
□ Capítulo 10

10.1.
a) 51 % no sorprendería pero 37% sí.
b) Ninguno de los dos porcentajes sorprendería si el tamaño de muestra fuera 35.

10.2.
i. (c) ii. (b) iii. (a) iv. (d)

10.3. (c) preciso y sin sesgo, a) preciso y con sesgo d) impreciso y sin sesgo, b) impreciso y con sesgo.

10.4. A continuación presentamos como ejemplo la cantidad de mujeres de cada muestra, al seleccionar 25 muestras de tamaño 3 de un curso que tiene 15 mujeres y 20 varones:



1 1 1 2 1 0 2 2 1 0 2 1 1 3 1 2 1 2 1 0 2 3 1 0 1 con un promedio de 1,28 mujeres por muestra de tamaño 3. Tenemos 4 muestras con 0 mujeres, 12 muestras con 1 mujer, 7 muestras con 2 mujeres y 2 muestras con 3 mujeres. Por azar se pueden obtener 4 entre 25 muestras con 0 mujeres. Por lo tanto, si la muestra seleccionada no tiene mujeres, no hay razones para sospechar discriminación.

10.5.

a) El margen de error para una confianza del 95% es aproximadamente $\frac{1}{\sqrt{n}}$

donde n es el tamaño de la muestra. La diferencia se debe a haber seleccionado distinta cantidad de varones y mujeres para realizar las entrevistas.

b) Es necesario incluir el margen de error debido a la variabilidad entre muestra y muestra.

□ Capítulo 13

13.1. Las tres respuestas son correctas. a) y b) son parte de la definición de estudios experimentales y observacionales. Una encuesta no impone ningún tratamiento, simplemente cuenta la frecuencia con la que aparece una respuesta, por lo tanto es observacional.

13.2. a) y b) son correctas y c) no. En un experimento tanto el grupo tratado como el grupo control son elegidos por el investigador. Es mejor cuando estos grupos son seleccionados al azar.

13.3. b) es el correcto. El primer estudio es observacional porque los sujetos no fueron asignados por el investigador al tratamiento (realización de actividad física).

13.4. a) es el correcto. El primer estudio es un estudio experimental con dos grupos tratamiento (1 litro; 2 litros) sin grupo control. El segundo es un estudio observacional porque el investigador no fijó cuanta leche se compraba en la casa para cada niño.

13.5 e) es el correcto. Se realizó un experimento en el cual los investigadores dividieron a los sujetos en grupo tratamiento y grupo control. Un censo debería estudiar a todos los que sufren de dolores de cabeza y no sólo a 20. Fueron comparadas las respuestas del grupo tratamiento (recibió chocolate) con la del grupo control (recibió un placebo). Las tabletas con gusto a menta sin chocolate eran el placebo.

13.6. a) Censo. Consulte en la dirección de la escuela donde podrán darle esa información.

Capítulo 15

15.1. Ambas preguntas se refieren a recodar nombres, esto no es una capacidad vinculada con la resolución de problemas, por lo tanto no son válidas para medir la inteligencia.

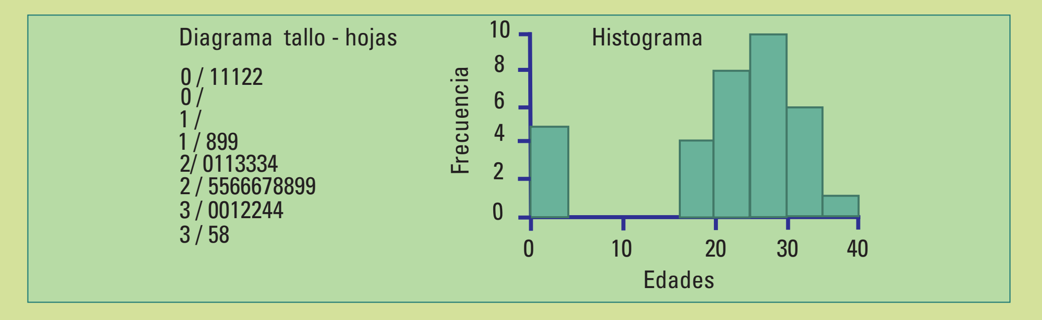
15.2. Comparar tasas de muerte, es decir la cantidad de muertes dividido la cantidad total de chicos que viaja en cada uno de los medios de transporte.

15.3.
a) La cantidad de muertes por cáncer aumenta a medida que la población se vuelve mayor. b) El porcentaje de muertes por cáncer aumenta a medida que la salud de la población mejora en general y no se muere por otras causas. c) El tiempo de supervivencia podría aumentar porque la enfermedad fue detectada antes; es decir porque el método diagnóstico (y no el tratamiento) es más efectivo.

15.4. Una manera claramente inválida de medir "estado físico" sería preguntar si le interesa o no le interesa la política. Una forma válida consiste en comparar las pulsaciones por minuto de la persona en posición acostada y luego parada, (manteniendo 5 minutos de reposo en cada una de las posiciones antes de realizar las mediciones). Cuanto menor sea la diferencia mejor será su estado físico.

Capítulo 17

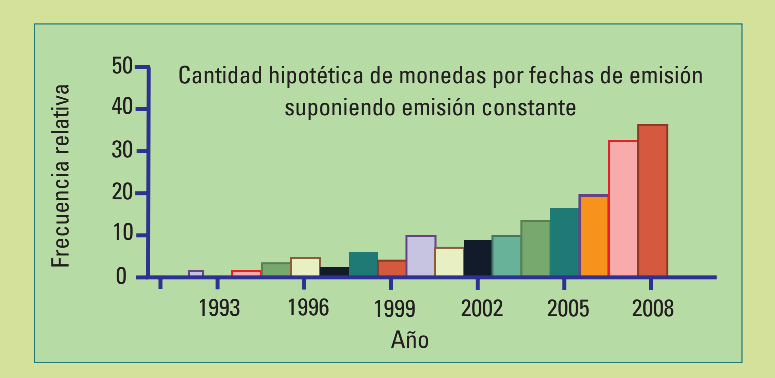
17.1. Tanto el diagrama tallo-hojas como el histograma muestran dos grupos de edades. El grupo de edades más pequeñas corresponde a la de los hijos, se trata de valores concentrados en el primer intervalo de clase. Las edades de los padres y madres se encuentran distribuidas en 5 intervalos de clase.



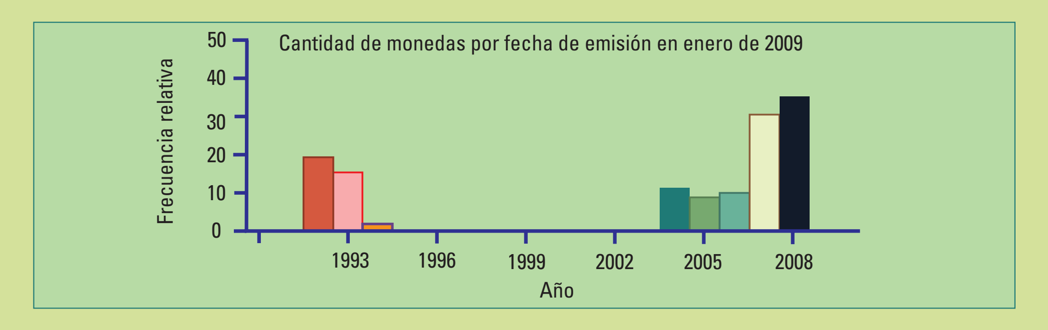
17.2. La mayoría de las familias no tendrán hijos y algunas tendrán hijos pequeños; aparecerá un pequeño promontorio en el lado más bajo de edades además del promontorio mayor correspondiente a las edades de los padres (c). Esta forma es similar a la observada en el diagrama tallo-hoja y el histograma del ejercicio 1.

17.3. Si la emisión de monedas fuera la misma todos los años y se perdieran una cierta proporción por año, se espera encontrar pocas monedas viejas y muchas nuevas. En este caso la distribución de las fechas de emisión tendría cola pesada a izquierda como se muestra en el siguiente histograma.

Año	Frecuencia	Frecuencia relativa
1992	20	0,15
1993	17	0,12
1994	1	0,01
2004	11	0,08
2005	9	0,07
2006	10	0,07
2007	33	0,24
2008	37	0,27



Utilizando 138 monedas de 10 centavos, seleccionadas en enero de 2008 obtenemos la siguiente distribución de frecuencias de la fecha de emisión:



El histograma resultante tiene una forma diferente a la esperada. Se observan dos grupos de fechas de emisión bien separados. El primero corresponde a las monedas emitidas entre los años 1992 y 1994, el segundo desde 2004 hasta 2008 estas últimas con una distribución más parecida a lo esperado. No encontramos monedas con año de emisión entre 1995 y 2003. Este resultado nos lleva a plantear nuevas preguntas, tal como suele ocurrir al realizar un análisis estadístico.

17.4. Para estudiar las longitudes de las palabras, seleccione un artículo de una revista de deportes y otro de una de divulgación científica. Para cada uno de los artículos obtenga:

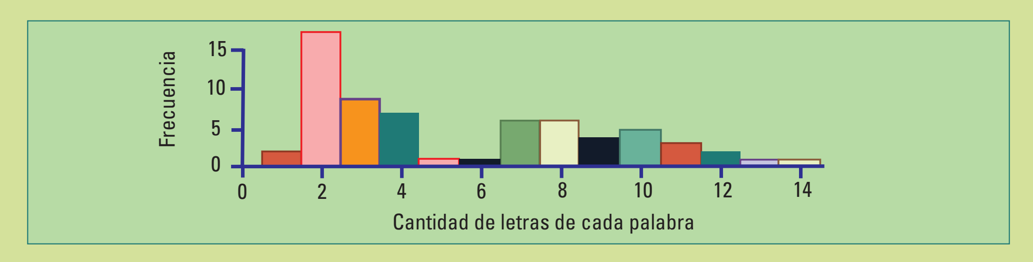
- la distribución de frecuencias
- la distribución de frecuencias relativas
- el histograma

de la variable "cantidad de letras" que tiene cada palabra. Compare las distribuciones obtenidas.

Lo haremos a continuación, como ejemplo, con el enunciado de este ejercicio:

Cantidad de letras por palabra	Palabras	Cantidad de palabras (Distribución de frecuencias)	Distribución de frecuencias relativas
1	y, a	2	0,03
2	de, un, de, de, de, de, de, la, la, de, de, el, de, la, de, lo, el, de	18	0,27
3	las, las, una, una, uno, los, que, las, con	9	0,14
4	Para, otro, Para, cada, cada, como, este	7	0,11
5	tiene	1	0,02
6	letras	1	0,02
7	revista, obtenga, palabra, Compare, haremos, ejemplo	6	0,09
8	estudiar, palabras, artículo, deportes, variable, cantidad	6	0,09
9	artículos, relativas, obtenidas, enunciado	4	0,06
10	longitudes, seleccione, científica, histograma, ejercicio	5	0,08
11	divulgación, frecuencias, frecuencias	3	0,05
12	distribución, distribución,	2	0,03
13	continuación	1	0,02
14	distribuciones	1	0,02
Total		66	1,03

Las frecuencias relativas no suman 1 por errores de redondeo.



Como la "cantidad de letras" es una variable numérica discreta los intervalos de clase del histograma están centrados en cada uno de los valores de la variable. El histograma muestra que el enunciado de este ejercicio tiene dos grupos de palabras, uno está formado por palabras cortas y el otro con palabras largas: el 55% son de a lo sumo 4 letras, el 40 % tiene entre 7 y 12 letras.

Un análisis similar puede realizarse con diferentes tipos de textos y en diferentes idiomas.

□ Capítulo 18

18.1.

e) Los diagramas tallo-hoja y los histogramas pueden mostrar los detalles que se esconden al calcular medidas resumen.

18.2.

d) La media, el desvío estándar, el máximo menos el mínimo, todos se ven afectados por la presencia de datos atípicos; la mediana y la distancia intercuartil, no.

18.3.

f) El desvío estándar sólo puede ser cero si todos los datos son iguales.

18.4.

b) Al dividir a todos los datos por 2, quedará el 20% entre 5 y 20. Luego al sumarles 10 quedará un 20% entre 15 y 30.

18.5) y 18.6)

a. Los **pesos** de varones y mujeres **por separado** tendrán medias y medianas similares si no hay alumnos o alumnas con pesos atípicos (es decir con pesos extremadamente bajos o extremadamente altos). Estos datos atípicos se detectan utilizando los gráficos - caja (box-plots). Si hay datos atípicos el desvío estándar puede ser demasiado grande y no representar a la mayoría de los pesos. Para los **pesos** de varones y mujere**s juntos,** las medidas resumen (media o mediana; distancia intercuartil o desvío estándar) pueden no ser una buena representación de los datos, porque se trata de dos grupos. En consecuencia el gráfico caja, que se obtiene a partir de medidas resumen, podría no dar una buena representación de los datos.

b. Registre la edad correspondiente a una fecha determinada. Por ejemplo el día que comienza el estudio. Utilice el formato de **edad decimal** expresada en años: Fecha del estudio: 29 de junio de 2008, Fecha de nacimiento: 15 de enero de 1992.

	Año	Mes	Día
Fecha del Estudio	2008	6	29
Fecha de nacimiento	1992	1	15
Edad	16	5	14

Edad: 16 años 5 meses y 14 días. Debemos convertir 5 meses y 14 días a años. Utilizaremos una forma sencilla que no tiene en cuenta la diferencia de días de los meses y los años bisiestos:

$$\text{Edad} = 16 + \frac{5}{12} + \frac{14}{365}$$

$$= 16{,}455 \text{ años.}$$

No deberían observarse demasiadas diferencias en las distribuciones de las edades de todos juntos y separados, varones y mujeres, como ocurre con peso. La media y el desvío estándar pueden resultar adecuadas.

18.7. Compare los resultados de los distintos años utilizando las medidas resumen que resultaron las más adecuadas para los pesos y para las edades y también histogramas.

18.8. La encuesta que realice en su división no será representativa de las opiniones de los demás años. Alumnos de años superiores suelen tener opiniones diferentes que los más chicos.

18.9. Para obtener **una muestra representativa de todos los años y de género**, utilice los registros de la dirección. Tome listados de varones y mujeres por separado. Puede utilizar un muestreo sistemático eligiendo, por ejemplo, uno de cada 10 alumnos/as.

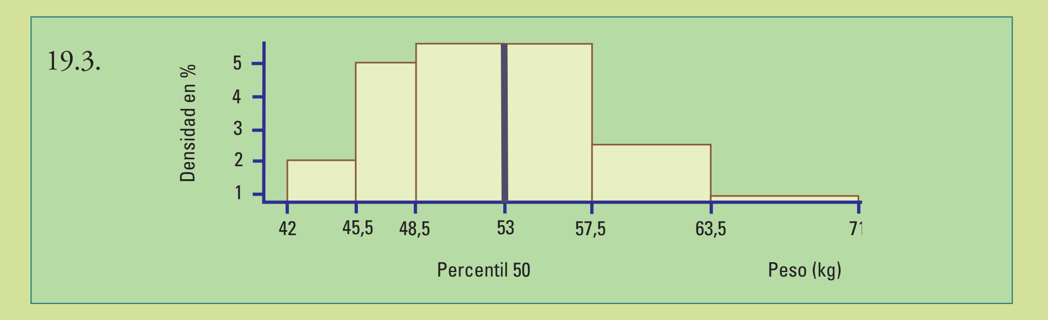
□ Capítulo 19

19.1. La figura 19.4 es una representación gráfica de los percentiles de los pesos. El percentil más pequeño, de 42 kg, es el del 3% indicando que hay un 3% de todas las mujeres de 16 años que tienen un peso menor. No sabemos cuánto menor, no sabemos cómo se distribuye ese 3% por debajo de 42 kg. Lo mismo ocurre con los pesos más altos que 71 kg, un 3% de los pesos de la población de mujeres de 16 años es mayor a 71 kg.

19.2. Complete la tabla de frecuencias siguiente utilizando la información de la figura 19.4

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %	Densidad Frecuencia Relativa en % / Longitud del Intervalo
[42 ; 45,5)	3,5	10 - 3 = 7	7 / 3,5 = 2,000
[45,5 ; 48,5)	3,0	25 - 10 = 15	15 / 3 = 5,000
[48,5 ; 53)	4,5	50 - 25 = 25	25 / 4,5 = 5,555
[53 ; 57,5)	4,5	75 - 50 = 25	25 / 4,5 = 5,555
[57,5 ; 63,5)	6,0	90 - 75 = 15	15 / 6 = 2,500
[63,5 ; 71)	7,5	97 - 90 = 7	7 / 7,5 = 0,933

19.3.



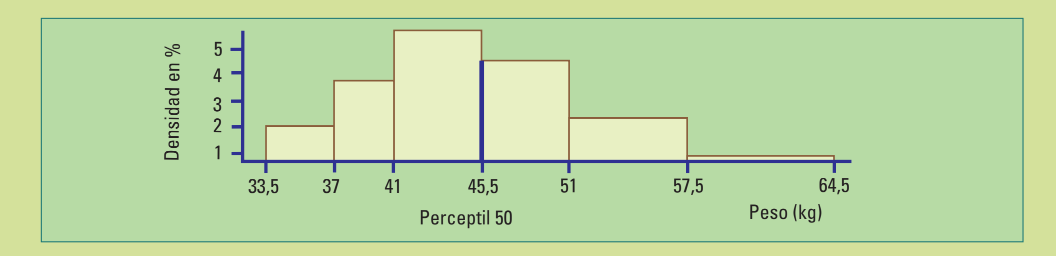
El área del histograma es la suma de las áreas de cada uno de los rectángulos de clase. Da como resultado 94 como se muestra a continuación:

Longitud de la base	Densidad (altura)	Área = Long. de la base x altura
3,5	7 / 3,5 = 2,000	3,5 x 7 / 3,5 = 7
3,0	15 / 3 = 5,000	3 x 15 / 3 = 15
4,5	25 / 4,5 = 5,555	4,5 x 25 / 4,5 = 25
4,5	25 / 4,5 = 5,555	4,5 x 25 / 4,5 = 25
6,0	15 / 6 = 2,500	6 x 15 / 6 = 15
7,5	7 / 7,5 = 0,933	7,5 x 7 / 7,5 = 7
Total		94

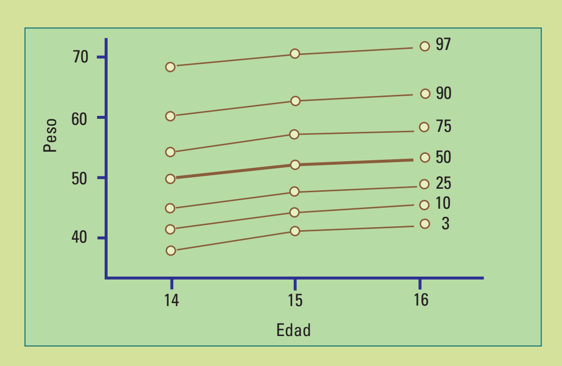
19.4. Percentiles del peso (kg) 13 años (mujeres)

Percentil	10	25	50	75	90	97
Peso	37	41	45,5	51	57,5	64,5

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %	Densidad Frecuencia Relativa en % / Longitud del Intervalo
[33,5 ; 37)	3,5	10-3 = 7	7 / 3,5 = 2,0000
[37 ; 41)	4,0	25-10 = 15	15 / 4 = 3,7500
[41 ; 45,5)	4,5	50 - 25 = 25	25 / 4,5 = 5,5550
[45,5 ; 51)	5,5	75-50 = 25	25 / 5,5 = 4,5454
[51 ; 57,5)	6,5	90-75 = 15	15 / 6,5 = 2,3100
[57,5 ; 64,5)	7 ,0	97-90 = 7	7 / 7 = 1,0000



19.5. y 19.6. La construcción de los diagramas tallo hoja permite ordenar los datos, esto facilita la obtención de los percentiles. Construya una tabla como la Tabla 19.2 y siga los pasos descriptos en la sección 19.1.2.



19.7. y 19.8. Debería obtener figuras similares a la siguiente:

En general, cada uno de los percentiles deberían aumentar con la edad; pero como se trata de diferentes alumnos en diferentes divisiones, pueden aparecer situaciones excepcionales y, entonces, no se cumpla ese aumento.

□ Capítulo 22

22.1.

a) Desde -1 hasta 1.

b) Cualquier número positivo.

22.2.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{1}{n-1}\,\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)\ \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}}$$

$$= \frac{1}{n-1}\,\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

22.3.

d) El coeficiente de correlación no cambia al sumar el mismo número a una de las variables o al multiplicarla por el mismo número.

22.4.

c) El coeficiente de correlación positivo muestra una tendencia de los valores mayores de una variable a estar acompañados de los valores mayores de la otra, pero para dos puntos en particular del diagrama todo es posible.

22.5.

c) Los tres diagramas muestran patrones no lineales fuertes. Sin embargo el coeficiente de correlación mide el grado de asociación lineal. Los dos primeros tienen r = 0 y el tercero un valor cercano a -1.

22.6.

b) Como el punto (3,11) pertenece a la recta y = 2 + b x, 11 = 2 + b x 3 entonces b= 3. Luego □ = 2 + 3.

22.7.

e) El coeficiente de correlación no puede ser mayor a 1.

22.8. El coeficiente de correlación se calcula como

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Vemos que las x's y las y's son intercambiables en la expresión anterior; no importa a los valores de cual de las variables llamemos x y a cual y.

Veamos ahora que cambiar de unidades producirá el mismo efecto en el numerador y el denominador y el resultado será el mismo:

Un cambio de unidades tiene la forma $z \rightarrow c\,z + d$. Si realizamos este cambio de unidades tendremos:

$$\frac{1}{n}\sum_{i=1}^{n}(cx_i+d) = c\,\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}d = c\,\bar{x}+d \qquad \frac{1}{n}\sum_{i=1}^{n}(cy_i+d) = c\,\frac{1}{n}\sum_{i=1}^{n}y_i + \frac{1}{n}\sum_{i=1}^{n}d = c\,\bar{y}+d$$

Por lo tanto el coeficiente de correlación en las nuevas unidades es igual al de las unidades originales:

$$r = \frac{\sum_{i=1}^{n}((cx_i+d)-(c\bar{x}+d))((cy_i+d)-(c\bar{y}+d))}{\sqrt{\sum_{i=1}^{n}((cx_i+d)-(c\bar{x}+d))^2\sum_{i=1}^{n}((cy_i+d)-(c\bar{y}+d))^2}} = \frac{c^2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{c^2\sum_{i=1}^{n}(x_i-\bar{x})^2 c^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Finalmente como la media y el desvío son influidas fuertemente por un valor atípico, el coeficiente de correlación también.

22.9. Siga los lineamientos del ejemplo 1 de este capítulo. Elija al azar, por ejemplo 10 alumnos de cada año y solicite que registren durante un mes la cantidad de horas que dedican cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promedien. Obtenga medidas resumen como las de la tabla 22.2 y gráficos para comparar la distribución de las horas y la edad entre varones y mujeres. Construya diagramas de dispersión de las horas y la edad para varones y mujeres por separado y evalúe si hay diferencias. Ajuste una recta de cuadrados mínimos en cada uno de los diagramas y obtenga el coeficiente de correlación. Estime la diferencia en la cantidad de horas dedicadas a actividades sedentarias, a cada edad, entre varones y mujeres como se realizó en el ejemplo 9 de este capítulo.

Capítulo 23

23.1.

a) Cuanto mayor sea el tamaño de la muestra menor será el desvío estándar de la distribución de muestreo.

b. Es válido en general, para poblaciones grandes y muestras pequeñas en comparación con el tamaño de la población.

c. Vale siempre.

d. Es el resultado del TCL.

23.2.

a) Verdadera

b) Falsa. La distribución de muestreo de $\hat{p}$ tiene un desvío estándar igual a $\sqrt{\frac{p(1-p)}{n}}$.

c. Falsa. La distribución de muestreo de $\hat{p}$ es aproximadamente Normal cuando np o n (1-p) son suficientemente grandes (se suele tomar 5 ó 10).

23.3. Éxito = salió 1 ó 2 al arrojar un dado.

a) Se arroja el dado n = 3 veces, se repite 60 veces y se obtienen los siguientes resultados:

Repetición nº	Cantidad de éxitos	$\hat{p}$
1	2	0,67
2	1	0,33
3	0	0,00
4	1	0,33
5	2	0,67
6	1	0,33
7	0	0,00
8	2	0,67
9	1	0,33
10	0	0,00
11	2	0,67
12	1	0,33
13	0	0,00
14	0	0,00
15	0	0,00
16	1	0,33
17	0	0,00
18	0	0,00
19	0	0,00
20	1	0,33

Repetición nº	Cantidad de éxitos	$\hat{p}$
21	0	0,00
22	1	0,33
23	0	0,00
24	0	0,00
25	1	0,33
26	0	0,00
27	2	0,67
28	2	0,67
29	2	0,67
30	1	0,33
31	1	0,33
32	1	0,33
33	2	0,67
34	3	1,00
35	2	0,67
36	1	0,33
37	1	0,33
38	0	0,00
39	0	0,00
40	0	0,00

Repetición nº	Cantidad de éxitos	$\hat{p}$
41	1	0,33
42	1	0,33
43	0	0,00
44	1	0,33
45	1	0,33
46	0	0,00
47	1	0,33
48	2	0,67
49	1	0,33
50	2	0,67
51	0	0,00
52	2	0,67
53	2	0,67
54	1	0,33
55	0	0,00
56	2	0,67
57	1	0,33
58	1	0,33
59	2	0,67
60	1	0,33

Al arrojar el dado 3 veces puede haber: ningún éxito, 1 éxito, 2 éxitos y 3 éxitos. Por lo tanto en la columna cantidad de éxitos solamente aparecen cuatro números (0; 1; 2 y 3) y también son cuatro los valores de la proporción de éxitos (0; 0,33; 0,77; 1)

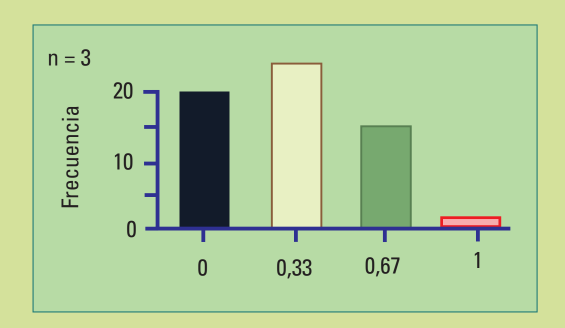
b) Se arroja el dado n = 30 veces, se repite 60 veces y se obtienen los siguientes resultados:

Repetición nº	Cantidad de éxitos	$\hat{p}$
1	10	0,33
2	11	0,37
3	12	0,40
4	10	0,33
5	10	0,33
6	10	0,33
7	13	0,43
8	14	0,47
9	4	0,13
10	10	0,33
11	16	0,53
12	3	0,10
13	9	0,30
14	9	0,30
15	13	0,43
16	10	0,33
17	11	0,37
18	9	0,30
19	8	0,27
20	14	0,47

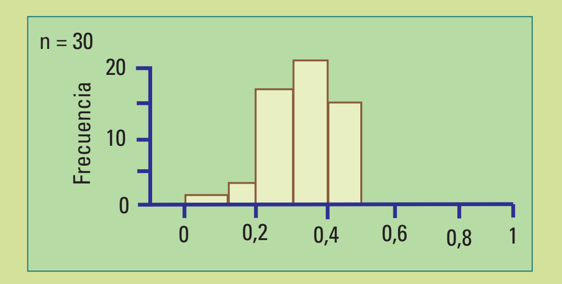
Repetición nº	Cantidad de éxitos	$\hat{p}$
21	10	0,33
22	13	0,43
23	13	0,43
24	9	0,30
25	12	0,40
26	8	0,27
27	5	0,17
28	13	0,43
29	10	0,33
30	9	0,30
31	8	0,27
32	14	0,47
33	7	0,23
34	10	0,33
35	8	0,27
36	4	0,13
37	13	0,43
38	14	0,47
39	14	0,47
40	8	0,27

Repetición nº	Cantidad de éxitos	$\hat{p}$
41	13	0,43
42	13	0,43
43	9	0,30
44	9	0,30
45	11	0,37
46	14	0,47
47	8	0,27
48	10	0,33
49	7	0,23
50	10	0,33
51	7	0,23
52	13	0,43
53	11	0,37
54	6	0,20
55	12	0,40
56	11	0,37
57	12	0,40
58	11	0,37
59	10	0,33
60	8	0,27

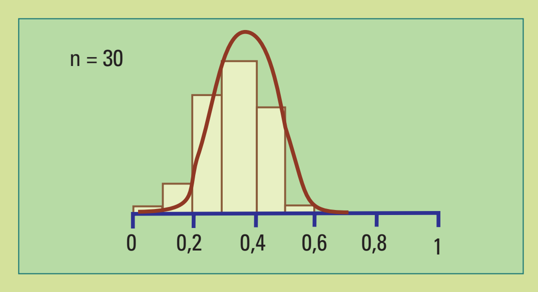
Al arrojar el dado 30 veces puede haber desde 0 hasta 30 éxitos. Hay también 31 valores diferentes en la columna encabezada por (0/30; 1/30; 2/30; ...;30/30)



El histograma anterior muestra la distribución de 60 proporciones muestrales de "éxito" en la repetición del experimento (arrojar un dado 3 veces, éxito= sale 1ó 2). El intervalo de clase que contiene al 0,33 es el más frecuente, pero la diferencia con las frecuencias del 0 y del 0,67 no es muy grande. Esta estimación está equivocada la mayoría de las veces.



El histograma anterior muestra la distribución de 60 proporciones muestrales de "éxito" en la repetición del experimento (arrojar un dado 30 veces, éxito= sale 1 ó 2). Las proporciones muestrales están más concentradas alrededor de la verdadera proporción (p =1/3) y su distribución puede aproximarse por la curva Normal:



En la práctica, la distribución de los valores de una variable en la población no cambia si se quita de ella una proporción muy pequeña. Los experimentos anteriores representan los resultados de muestreos aleatorios simples para estimar la proporción de éxitos (p) en una población (que solamente nosotros sabemos tiene p =1/3). Se trata de una simulación, repetimos muchas veces el cálculo de un estimador para conocer su distribución de muestreo, pero cuando lo usamos como estimador lo hacemos una única vez. Los histogramas anteriores muestran que es mejor usar muestras de tamaño 30 que de tamaño 3 para estimar una proporción. Cuanto más grande, mejor, siempre que el procedimiento de selección no tenga sesgo.

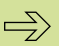
23.4.

a) Una calculadora da μ = 8,5 y σ = 5,346338. Se trata de parámetros poblacionales. N = 6 es el tamaño de la población. La fórmula de cálculo del desvío estándar poblacional difiere de s en que tiene $\sqrt{N}$ en vez de $\sqrt{(n-1)}$ en el denominador:

$$\sigma_X = \sqrt{\frac{1}{6}\sum_{i=1}^{6}(x_i - \mu_X)^2}$$

b) Enumere los valores de la variable de todas las 15 posibles muestras de tamaño n =2 y calcule $\bar{x}$ para cada una de ellas.

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %
Muestra 1	(2, 4)	$\bar{x}_1 = 3$
Muestra 2	(2,6)	$\bar{x}_2 = 4$
Muestra 3	(2, 9)	$\bar{x}_3 = 4,5$
Muestra 4	(2, 12)	$\bar{x}_4 = 7$
Muestra 5	(2, 18)	$\bar{x}_5 = 10$
Muestra 6	(4,6)	$\bar{x}_6 = 5$
Muestra 7	(4,9)	$\bar{x}_7 = 6,5$
Muestra 8	(4,12)	$\bar{x}_8 = 8$
Muestra 9	(4,18)	$\bar{x}_9 = 11$



Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %
Muestra 10	(6,9)	$\bar{x}_{10} = 7{,}5$
Muestra 11	(6,12)	$\bar{x}_{11} = 9$
Muestra 12	(6,18)	$\bar{x}_{12} = 12$
Muestra 13	(9,12)	$\bar{x}_{13} = 10{,}5$
Muestra 14	(9,18)	$\bar{x}_{14} = 13{,}5$
Muestra 15	(12,18)	$\bar{x}_{15} = 15$

Observación. En este ejercicio se conoce completamente la población, y la distribución de muestreo de los valores de la media muestra.

c) Muestre que la media de las 15 medias muestrales ($\bar{x}$) es μ.

La media de las 15 medias muestrales ($\bar{x}$) es un parámetro poblacional ($\mu_{\bar{x}}$). La población (N = 15) está formada por los 15 valores posibles de $\bar{x}$.

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{n} \bar{x}_i}{N}$$

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{15} \bar{x}_i}{15}$$

$$= \frac{3 + 4 + 5{,}5 + 7 + 10 + 5 + 6{,}5 + 8 + 11 + 7{,}5 + 9 + 12 + 10{,}5 + 13{,}5 + 15}{15}$$

$$= \frac{127{,}5}{15}$$

$$= 8{,}5$$

$$\mu_{\bar{x}} = \mu$$

d) El desvío estándar de las 15 medias muestrales ($\bar{x}$) es un parámetro poblacional ($\sigma_{\bar{x}}$) de la misma población considerada en **c.** Usando una calculadora obtenemos el desvío estándar de los 15 valores:

$$\sigma_{\bar{x}} = \frac{1}{15} \sum_{i=1}^{15} (\bar{x}_i - \mu_{\bar{x}})^2$$

$$= 3{,}381321$$

El desvío estándar de las 15 medias muestrales es: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

En efecto:

n = 2, N = 6, σ = 5,346338 $\quad \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{5{,}346338}{\sqrt{2}} \sqrt{\frac{6-2}{6-1}}$

$$= 3{,}381321$$

$$= \sigma_{\bar{x}}$$

e) En general, cuando el tamaño de la población N es muy grande en comparación con el tamaño muestral n, $\frac{N-n}{N-1} \approx 1$ y la expresión anterior se simplifica a $\frac{\sigma}{\sqrt{n}}$.

El error estándar no depende, en este caso, del tamaño de la población.

□ Capítulo 24

24.1.

a) Como el tamaño de la muestra n = 40 es mayor a 30, podemos utilizar la aproximación por la Normal a la distribución de la media muestral. En este caso, el intervalo aproximadamente del 95% de confianza para la media muestral es de la forma $\bar{x} \pm 2 \frac{\sigma}{\sqrt{n}}$ por lo tanto resulta $3{,}09 \pm 2 \cdot \frac{0{,}25}{\sqrt{40}} \Rightarrow 3{,}09 \pm 0{,}08$

El intervalo es [3,01; 3,17] y tiene longitud 0,16.

b) $3{,}09 \pm 3 \cdot \frac{0{,}25}{\sqrt{40}} \Rightarrow 3{,}09 \pm 0{,}12$, es el resultando el intervalo [2,97; 3,21] que tiene longitud 0, 24.

c) La longitud del intervalo de confianza del 95% es $4 \frac{\sigma}{\sqrt{n}}$. Debemos hallar n tal que $4 \cdot \frac{0{,}25}{\sqrt{n}} = 0{,}1$. Por lo tanto $n = \left(4 \frac{0{,}25}{0{,}1}\right)^2$

$n = 100$

24.2.

c) es la correcta.

Las longitudes de los intervalos de confianza son para 95% y 99,7% respectivamente:

por el procedimiento 1 $4\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ y $6\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

por el procedimiento 2 $4\sqrt{\frac{1}{n}}$ y $6\sqrt{\frac{1}{n}}$

Para cualquiera de los dos procedimientos:

$$\frac{100 \text{ x (Longitud para 99,7\% - Longitud para 95\%)}}{\text{Longitud para 95\%}} = \frac{100 \text{ x } (6-4)}{4}$$
$$= \frac{100 \text{ x} 1}{2}$$
$$= 50$$

24.3.

c) es la correcta. El desvío estándar se estima por s = 1,4 y el intervalo de confianza es de la forma $\left[\bar{x}-2\left(\frac{s}{\sqrt{n}}\right);\bar{x}+2\left(\frac{s}{\sqrt{n}}\right)\right]$, quedando

$\left[6{,}8-2\ \frac{1{,}4}{\sqrt{49}}\ ;\ 6{,}8+2\ \frac{1{,}4}{\sqrt{49}}\right]$. Esto es [6,8 -0,4; 6,8 +0,4]

24.4. No hay garantías que la verdadera proporción se encuentre dentro del intervalo. Ninguna de las respuestas es correcta.

24.5.

d) es correcta. Al aumentar el tamaño de la muestra de **n** a **kn**, la longitud del intervalo se divide por $\sqrt{\mathbf{k}}$**.**

24.6.

e) es correcta. Cuando se construye un intervalo de confianza se confía que el verdadero valor se encuentra dentro de él, pero no se puede estar seguro/a.

24.7.

b) es correcta. $n_1 = 400$ $n_2 = 300$ $\hat{p}_1 = 0{,}65$ $\hat{p}_2 = 0{,}48$

El desvío estándar estimado de la diferencia de proporciones es:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0{,}65(1-0{,}65)}{400}+\frac{0{,}48(1-0{,}48)}{300}}$$

$$= 0{,}0374$$

el IC para $\hat{p}$ - $\hat{p}$ es: (0,65-0,48) ± (2 x 0,0374) = 0,17 ± 0,0748

24.8. Primero verificamos que n $\hat{p}$ = 16 > 10 y n (1- $\hat{p}$) = 220 - 16 = 204 > 10. Sabemos que 220 es el 0,44% < 10 % de la población por lo que no es necesario utilizar una corrección por el tamaño de la población (el factor de corrección sería 0,998). La proporción muestral de piezas dañadas es $\hat{p}$ = 16 /220 = 0,072 y el desvío estándar estimado de $\hat{p}$ es:

$$\sigma_{\hat{p}} \text{ estimado} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \sqrt{\frac{0{,}072(1-0{,}072)}{220}}$$

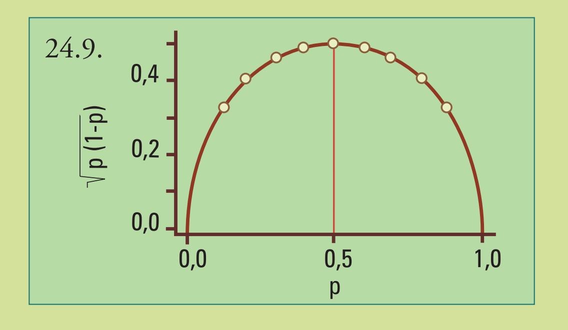
$$= 0{,}0174$$

El intervalo del 95% de confianza para la verdadera proporción de piezas dañadas es 0,072 ± 2 . 0,0174. Luego, tenemos un 95% de confianza que la proporción de piezas dañadas se encuentra entre 0,0372 y 1,1068. Como la cantidad de piezas del embarque es 50.000 este intervalo se traduce en el siguiente intervalo de piezas dañadas:

$$0{,}0372 \times 50.000 = 1.860 \quad y \quad 1{,}1068 \times 50.000 = 5.340$$

Tenemos un 95% de confianza que la cantidad de piezas dañadas se encuentra entre 0,0372 y 1,1068.

24.9.



El gráfico de $\sqrt{p(1-p)}$ en función de p alcanza su máximo para p = 0,5.

☐ Capítulo 25

25.1. Cuando se rechaza la hipótesis nula con un valor -p = 0,03 eso significa que

a. Falsa. Es posible rechazar la hipótesis nula en forma equivocada
b. Falsa. El valor-p nada dice respecto a la proporción de veces que la hipótesis nula es falsa.
c. Falsa.
d. Verdadera. El valor -p se calcula suponiendo que la hipótesis nula es verdadera.

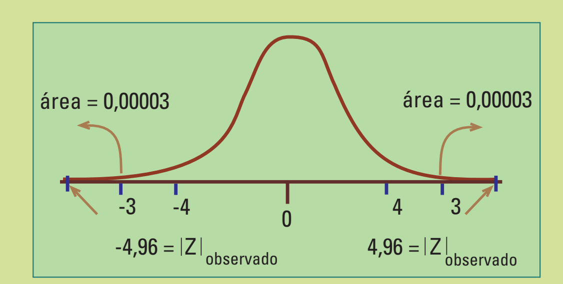
25.2.
a. Verdadera. Es necesario tener razones vinculadas a la naturaleza del problema a estudiar y no con los datos, para elegir una hipótesis alternativa unilateral.
b. Falsa. La elección de la hipótesis alternativa no debe basarse en los datos.
c. Falsa. La elección de la hipótesis alternativa no depende de la magnitud del efecto que se interesa probar.

25.3.
a. Falsa. La elección de la hipótesis alternativa no debe basarse en los datos.
b. Verdadera. Es necesario tener razones relacionadas con la naturaleza del problema a estudiar para elegir una hipótesis alternativa unilateral
c. Falsa. La elección de la hipótesis alternativa no está relacionada con el cálculo del valor - p.
d. Falsa. La elección de la hipótesis alternativa no depende de la cantidad de datos.
d. Falsa. La elección de la hipótesis alternativa no depende de la cantidad de datos.

25.4.
e. Verdadera. Aunque tenga la sospecha respecto a la dirección del resultado (las ambulancias tardan más que el tiempo especificado en H_0) debe plantear una alternativa bilateral, a menos que pueda probar que es imposible que el tiempo sea menor a 12 minutos.

25.5.

a. Verdadera. Tenemos. H_0: μ =12 y Ha: μ ≠ 12.



$n = 44, \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{44}}$

$= \frac{8}{6,63}$

$= 1,21$

$z_{observado} = \frac{18 - 12}{1,21}$

$= 4,96$

El área hacia valores más alejados del cero desde -4,96 y 4,96 es menor al área 2 . 0,00003 =0,00006 de los valores más alejados desde -4 y 4.

Por lo tanto el valor -p (4,96; a dos colas) < 0,00006. Esto significa que si la hipótesis fuera verdadera, tendríamos un valor como el observado o más extremo menos de 1 de 1000 veces. Concluimos que la hipótesis nula no es verdadera y se rechaza la afirmación del intendente.

Un intervalo del 95% de confianza para el tiempo medio de tardanza de una ambulancia es más informativo:

$$\bar{x} \pm 1,96 \, . \, \frac{s}{\sqrt{n}} = 18 \pm 1,96 \, . \, 1,21$$

$$= 18 \pm 2,37$$

$$[\, 15,63 \, ; 20,37 \,]$$

Con un 95% de confianza afirmamos que el tiempo medio que tarda una ambulancia en llegar al lugar del accidente se encuentra entre 15,63 y 20,37 minutos.

25.6.

a. y b. Falsas. Los datos no son parte de las hipótesis.

c. Verdadera. Hay que realizar un test para diferencia de proporciones bilateral por que no hay razones para suponer que la diferencia tendrá necesa ria mente una dirección.

d. Falsa .

25.7.

c. Verdadera. Aunque tenga la sospecha respecto a la dirección del resultado (las ambulancias tardan más que el tiempo especificado en H_0) debe plantear una alternativa bilateral, a menos que pueda probar que es imposible que el tiempo sea menor a 12 minutos.

Tenemos H_0: $p_1 - p_2 = 0$ y Ha: $p_1 - p_2 \neq 0$

$\hat{p}_1 = \frac{68}{140}$ $\hat{p}_2 = \frac{95}{180}$

$\hat{p}_1 = 0,49$ $\hat{p}_2 = 0,53$

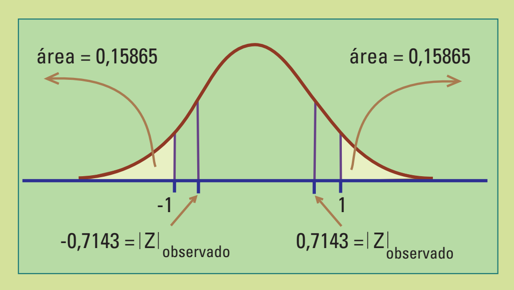
El estadístico del test es

$$\frac{\hat{p}_1 - \hat{p}_1 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

y $$z_{observado} = \frac{\frac{68}{140} - \frac{95}{180}}{\sqrt{\frac{\frac{68}{140}\left(1 - \frac{68}{140}\right)}{140} + \frac{\frac{95}{180}\left(1 - \frac{95}{180}\right)}{180}}}$$

$$= \frac{0{,}49 - 0{,}53}{0{,}056}$$

$$= -0{,}7143$$



Como el test es a dos colas el p-valor es la suma de las áreas de las 2 colas, hacia valores alejados del cero bajo la curva N(0,1) a izquierda de -0,7143 y a derecha de 0,7143. Estas **áreas son mayores** que 2.0,15865 (las áreas que dejan hacia las colas el -1 y el 1, figura anterior).

27.8. ¿Cómo va a elegir las muestras? ¿Serán representativas de todos los profesores y alumnos del país? Si no lo es, restrinja el alcance de sus conclusiones a la población a que considere adecuada. Por ejemplo a todos los profesores y alumnos de su provincia, de su ciudad o de su escuela. Verifique que el tamaño de las muestras es suficientemente grande como para utilizar el TCL para la distribución de muestreo de las proporciones muestrales.

Llame "éxito"= una persona piensa que el cambio climático va afectar fuertemente sus vidas en los próximos 20 años.

Realice un test para diferencia de proporciones como en el del ejercicio 7 para comparar las opiniones entre:

- profesores y alumnos, p_1= proporción poblacional de éxitos entre los/as profesores/as, p_2= proporción poblacional de éxitos entre los alumnos/as
- mujeres y varones, p_1= proporción poblacional de éxitos entre los alumnos, p_2= proporción poblacional de éxitos entre las alumnas

Construya una estimación de la verdadera diferencia de proporciones utilizando un intervalo de confianza del 95%.

27.9.

c. Falsa. Los valores-p chicos (menores que 0,05) son evidencia en contra de la hipótesis nula.
b. Falsa. El tamaño de muestra no corrige el sesgo.
c. Falsa.
d. Verdadera.
e. Falsa.

◻ Bibliografía recomendada

Estadística. David Friedman, Robert Pisani, Roger Purves y Ani Adhikari. **ISBN:** 848585568X 1993 Editor Antoni. Bosch Barcelona.

Interactive Statistics. Martha Aliaga, Brenda Gunderson. **ISBN:** 031497561 Prentice Hall 2006.

Introduction to the practice of Statistics. David S. Moore, George P McCabe. **ISBN:** 9780716764007. W.H. Freeman & Company 1989.

Statistics. Concepts and controversies. David S. Moore, William I. Notz **ISBN:** 9780716786368. W.H. Freeman & Company 2006